

**INVESTIGATING THE DIVERSITY AND EVOLUTION
OF TEMPERATE ACTINOBACTERIOPHAGES**

by

Travis Nathaniel Mavrich

Bachelor of Science, Pennsylvania State University, 2004

Submitted to the Graduate Faculty of
The Dietrich School of Arts and Sciences in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH
DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented by

Travis Nathaniel Mavrich

It was defended on

March 12, 2019

and approved by

Karen M. Arndt, PhD, Professor, Dept. of Biological Sciences

Jon P. Boyle, PhD, Associate Professor, Dept. of Biological Sciences

Jeffrey G. Lawrence, PhD, Professor, Dept. of Biological Sciences

Nathan L. Clark, PhD, Associate Professor, Dept. of Computational and Systems Biology

Roger W. Hendrix, PhD, Distinguished Professor, Dept. of Biological Sciences (in Memoriam)

Dissertation Advisor: Graham F. Hatfull, PhD, Professor, Dept. of Biological Sciences

Copyright © by Travis Nathaniel Mavrich

2019

INVESTIGATING THE DIVERSITY AND EVOLUTION OF TEMPERATE ACTINOBACTERIOPHAGES

Travis Nathaniel Mavrich, PhD

University of Pittsburgh, 2019

Bacteriophages (phages) are viruses that infect bacteria, and they have been evolving for billions of years to combat the hosts they infect as well as other phages in the environment. Many phages are temperate, and after infection they may remain in the host as a prophage, forming a lysogen instead of initiating lytic growth. Lysogeny presents unique challenges and benefits, and as a result temperate phages impact their environment differently than obligately lytic phages. Although temperate phages are diverse, many paradigms about their lifestyle and evolution have been derived from small collections of phages representing limited and ill-defined genetic diversity. Therefore, I have investigated a large collection of phages infecting hosts in the phylum Actinobacteria to expand our understanding of temperate phage diversity and evolution. First, I show that in contrast to obligately lytic phages, temperate phages evolve within two evolutionary modes that are characterized by different degrees of gene content flux and that vary by the type of temperate phage and its bacterial host. Second, I characterize extrachromosomal *Mycobacterium* prophages that utilize partitioning systems to maintain lysogeny, which are not commonly reported. I show that these prophages exhibit partition-mediated incompatibility and that components of this system evolve under different selective pressures possibly to avoid this instability. Third, I characterize *Bifidobacterium* prophages and explore how they interact with their hosts. Some of these prophages utilize a unique integration site and encode a shufflon system. This shufflon may control host range and is the first to be reported in actinobacteriophages. Last, I examine the evolution of a *Mycobacterium* prophage immunity

system. This regulatory circuit enables prophages to control expression of lytic genes, maintaining lysogeny and defending against superinfection. I show that closely-related phages with diverging immunity systems generate a complex immunity network and gaining virulence to escape this network is difficult. Overall, this research has broadened our understanding of temperate phage diversity and evolution, and it has identified genetic systems that can be used to develop new genetic tools.

TABLE OF CONTENTS

PREFACE.....	xxi
1.0 INTRODUCTION.....	1
1.1 BACTERIAL DIVERSITY	2
1.2 BACTERIAL TRANSCRIPTION.....	3
1.3 PHAGE EVOLUTION AND MOSAICISM.....	5
1.4 TEMPERATE PHAGES	8
1.4.1 Prevalence of lysogeny	9
1.4.2 The genetic switch	10
1.4.3 Evolution of temperate phages.....	12
1.5 IMMUNITY SYSTEMS	14
1.5.1 Enterobacteria phage λ	14
1.5.2 Enterobacteria phage P2	16
1.5.3 <i>Bacillus</i> phage Φ 105.....	16
1.5.4 <i>Mycobacterium</i> phage BPs	17
1.5.5 Miscellaneous λ -related systems	17
1.5.6 Enterobacteria phage P22	18
1.5.7 Enterobacteria phage P1	19
1.5.8 Enterobacteria phage N15.....	20
1.5.9 <i>Streptomyces</i> phage phiC31	21
1.5.10 <i>Mycobacterium</i> phage L5.....	21
1.5.11 Superinfection immunity	24

1.6 INHERITANCE SYSTEMS.....	27
1.6.1 Integration systems	27
1.6.2 Partitioning systems	28
1.7 IMPACT ON BACTERIAL HOST	30
1.8 BIOTECHNOLOGICAL APPLICATIONS	32
1.9 ACTINOBACTERIOPHAGE DIVERSITY	33
1.10 CURRENT CHALLENGES.....	35
2.0 INVESTIGATION OF WHOLE GENOME EVOLUTION	39
2.1 INTRODUCTION	39
2.2 MATERIALS AND METHODS.....	41
2.2.1 Phages used in this study	41
2.2.2 Collection of virus metadata	41
2.2.3 Categorization of phams into general gene functions.....	42
2.2.4 Prediction of phage lifestyle	43
2.2.5 Calculation of whole genome gene content dissimilarity	44
2.2.6 Calculation of whole genome nucleotide distance	44
2.2.7 Plotting genomic similarity.....	45
2.2.8 Assigning evolutionary mode	46
2.2.9 Calculation of genome size disparities	47
2.2.10 Calculation of shared and unshared gene subset data.....	47
2.2.11 Analysis of prokaryotic Virus Orthologous Groups	49
2.2.12 Creating cluster-specific multi-gene phylogenies	49
2.2.13 Measuring rates of horizontal gene transfer	51

2.2.14 Analysis of LysB horizontal gene transfer	52
2.2.15 Data analysis	52
2.2.16 Computing MaxGCDGap	52
2.3 RESULTS.....	53
2.3.1 Generation of genomic similarity plots	53
2.3.2 Phages exhibit two evolutionary modes	60
2.3.3 Genetically related phages exhibit specific evolutionary modes.....	67
2.3.4 Evolutionary modes are correlated with phage lifestyles.....	70
2.3.5 Evolutionary modes are correlated with HGT	74
2.3.6 Cluster A phages exhibit two evolutionary modes	77
2.3.7 Temperate HGCF phages exhibit greater rates of HGT	81
2.3.8 Evolutionary modes differ by host phyla	85
2.3.9 Implications of evolutionary modes.....	87
2.3.10 Quantification of phage genetic isolation using MaxGCDGap.....	89
2.4 DISCUSSION.....	93
3.0 CHARACTERIZATION AND INDUCTION OF BIFIDOPROPHAGES.....	97
3.1 INTRODUCTION	97
3.2 MATERIALS AND METHODS.....	99
3.2.1 Bacterial strains.....	99
3.2.2 Prophage characterization	101
3.2.3 Phamerator database construction.....	103
3.2.4 Optimization of mitomycin C induction	103
3.2.5 Bifidobacterial growth and mitomycin C induction	103

3.2.6 Induction verification using PCR	104
3.2.7 Plaque assays	104
3.2.8 Induced phage genome sequencing	105
3.2.9 Induced phage replication quantification	106
3.2.10 Transmission electron microscopy	106
3.2.11 Flow cytometry sample preparation and processing	107
3.2.12 Flow cytometry data analysis	108
3.2.13 Rin shufflon analysis	109
3.2.14 Rin shufflon analysis in WGS reads	110
3.2.15 <i>dnaJ₂</i> -integrating phage attachment site analysis	110
3.2.16 Host 16S rRNA analysis	111
3.2.17 Gene content flux analysis	111
3.3 RESULTS	111
3.3.1 Bioinformatic characterization of bifidophages	111
3.3.2 Prophages can be induced with mitomycin C	120
3.3.3 Phage genomes circularize after induction	122
3.3.4 Induced phages increase in copy number	124
3.3.5 Induced phages contain packaged DNA	127
3.3.6 Induction generates complete phage particles	131
3.3.7 Characterization of the Rin shufflon	133
3.3.8 tRNA ^{Met} -integrated prophages harbor an inversion locus	141
3.3.9 Analysis of bifidophage host ranges	142
3.3.10 Analysis of bifidophage evolutionary modes	145

3.4 DISCUSSION.....	146
4.0 CHARACTERIZATION OF PARTITIONING SYSTEMS	148
4.1 INTRODUCTION	148
4.2 MATERIALS AND METHODS.....	150
4.2.1 Phamerator database construction.....	150
4.2.2 Generation of phages and lysogens.....	150
4.2.3 RNAseq	151
4.2.4 DNAseq	151
4.2.5 Phylogenetic analysis of partitioning cassettes	152
4.2.6 Prediction of partitioning types	153
4.2.7 ParA and ParB coevolution analysis	153
4.2.8 Prediction of <i>parS-L</i> and <i>parS-R</i> loci.....	154
4.2.9 ParB Purification	155
4.2.10 Electrophoretic mobility shift assays.....	156
4.2.11 Construction of <i>parABS</i> plasmids and plasmid retention assay	157
4.2.12 DNA skew analysis	157
4.2.13 Partitioning cassette incompatibility assay	158
4.2.14 Testing for a RedRock <i>parABS</i> origin of replication	159
4.3 RESULTS.....	160
4.3.1 RedRock contains a partitioning cassette	160
4.3.2 Characterization of Cluster A partitioning systems	164
4.3.3 RedRock partitioning genes are expressed during lysogeny	168
4.3.4 Multiple copies of RedRock are maintained during lysogeny	170

4.3.5 RedRock ParB exhibits <i>parS</i> binding specificity.....	172
4.3.6 RedRock <i>parABS</i> increases plasmid retention	173
4.3.7 RedRock <i>parABS</i> confers replicon incompatibility	174
4.3.8 RedRock and Alma ParB homologs exhibit distinct specificities	176
4.3.9 <i>parABS</i> systems confer prophage incompatibility	178
4.3.10 Evolution of Cluster A <i>parABS</i> systems	180
4.3.11 Investigation of RedRock prophage origin of replication	182
4.3.12 Analysis of other partitioning systems	185
4.4 DISCUSSION.....	187
5.0 FUNCTION AND EVOLUTION OF CLUSTER A IMMUNITY SYSTEMS	189
5.1 INTRODUCTION	189
5.2 MATERIALS AND METHODS.....	191
5.2.1 Phamerator database construction.....	191
5.2.2 Identification and analysis of stoperator sequences.....	191
5.2.3 Computation of genomic similarity metrics	192
5.2.4 Genetic distance of specific Cluster A genes.....	193
5.2.5 Repressor nucleotide alignment and phylogeny.....	194
5.2.6 Preparation of phage lysates and lysogens	194
5.2.7 RNAseq	198
5.2.8 Repressor overexpression and EMSAs	198
5.2.9 Construction of cloned repressor strains	199
5.2.10 Engineering L5 derivatives with C-terminally tagged repressors.....	201
5.2.11 Western blot analysis of FLAG-tagged L5 repressor	202

5.2.12 Superinfection immunity assays	203
5.2.13 Isolation of defense escape mutants	204
5.2.14 Genomic analysis of D29 and its relatives.....	204
5.2.15 Construction of strains carrying repressor-controlled P _{left} locus	205
5.2.16 P _{left} toxicity test	207
5.2.17 Data analysis	207
5.3 RESULTS	208
5.3.1 Characterization of the Cluster A immunity system	208
5.3.2 L5 clade phages exhibit mesoimmunity	218
5.3.3 Mesoimmunity phenotypes correlate with immunity system evolution.....	226
5.3.4 Mesoimmunity is repressor-mediated	229
5.3.5 Characterization of defense escape mutants	232
5.3.6 An engineered L5 mutant exhibits acute homotypic virulence.....	240
5.3.7 Evolution of stoperators	246
5.3.8 Expression from P _{left} is toxic	253
5.3.9 An extended cloned repressor locus strengthens immunity	255
5.3.10 The un-regulated extended repressor locus is toxic	257
5.3.11 A Bxb1 DEM escapes a Bxb1 CRS	262
5.4 DISCUSSION.....	263
6.0 CONCLUSIONS	268
6.1 SUMMARY OF RESULTS	268
6.1.1 Whole genome evolutionary patterns.....	268
6.1.2 <i>Bifidobacterium</i> prophages	270

6.1.3 Partitioning systems	271
6.1.4 Immunity systems	272
6.2 FUTURE DIRECTIONS	274
6.2.1 Dynamics of lysogeny	274
6.2.2 Tracking host history	275
6.2.3 Tracking horizontal gene transfer	276
6.2.4 High throughput phage genomics.....	277
6.2.5 Biotechnological applications.....	277
6.2.6 Evolution of viral latency.....	278
6.3 CODA	279
APPENDIX A PHAMERATOR DATABASE MANAGEMENT.....	280
A.1 INTRODUCTION	280
A.2 PIPELINE DEVELOPMENT	284
A.2.1 Overview of SEA-PHAGES data pipeline	284
A.2.2 Overview of key aspects of PhameratorDB	286
A.2.3 Modifications to PhameratorDB and data management scripts	289
A.2.4 Development of ticket tracking systems	294
A.2.5 Retrieve new data to import into PhameratorDB.....	298
A.2.6 Import new data into PhameratorDB	302
A.2.7 Update specific fields.....	314
A.2.8 Identify conserved protein domains	316
A.2.9 Group all proteins into phamilies	317
A.2.10 Export updated database.....	318

A.2.11 Managing the record of tickets	320
A.2.12 Freeze database for publication	320
A.2.13 Compare phage databases for consistency.....	322
A.3 CONCLUSIONS.....	324
APPENDIX B OLIGONUCLEOTIDES USED IN THIS STUDY	326
APPENDIX C PLASMIDS USED IN THIS STUDY	329
BIBLIOGRAPHY	330

LIST OF TABLES

Table 2-1. Phams used to construct multi-gene phylogenies.....	50
Table 3-1. Bifidobacterium genomes used in this study.....	100
Table 3-2. Bifidoprohage genomes analyzed in this study.....	102
Table 3-3. Dimensions of bifidophages detected by TEM.	132
Table 3-4. Bifidoprohage <i>attL</i> and <i>attR</i> common core sites.	143
Table 4-1. Description of ParB EMSA substrates.	156
Table 4-2. Comparison of RedRock and L5 prophage copy number.	171
Table 4-3. RedRock ParB increases plasmid retention.....	174
Table 5-1. Phages used in this chapter.....	196
Table 5-2. Infection scoring strategy.	204
Table 5-3. D29 sub-clade single nucleotide polymorphisms.....	247
Table 5-4. D29 sub-clade insertions and deletions.	247
Table 5-5. D29 sub-clade stopoperator sites.....	248
Table A-1. Python scripts to maintain PhameratorDB.	293
Table A-2. Structure of import tickets.....	297
Table A-3. Required arguments for retrieve_database_updates.py script.	298
Table A-4. Required arguments for import_phage.py script.	302
Table A-5. Source of data used to populate <i>Phage</i> table.....	305
Table A-6. Source of data used to populate <i>Gene</i> table.....	307
Table A-7. Quality control options differ between run modes.	312
Table A-8. Structure of update_[field] tables.	315

Table A-9. Required arguments for update_[field].py script.....	315
Table A-10. Required arguments for cdd_pp.py script.	316
Table A-11. Required arguments for k_phamrate.py script.....	317
Table A-12. Required arguments for export_database.py script.	318
Table A-13. Required arguments for freeze_database.py script.....	321
Table C-14. Required arguments for compare_databases.py script.	323
Table C-15. Files generated from compare_databases.py script.	324
Table D-1. Oligonucleotides used in this study.0.....	326
Table E-1. Plasmids used in this study.0.....	329

LIST OF FIGURES

Figure 1-1. Diagram of temperate phage lifecycles.....	9
Figure 1-2. Genome map of <i>Mycobacterium</i> phage L5.....	23
Figure 1-3. Diversity across a phage genome landscape.....	36
Figure 2-1. Comparison of gene content derived from kmer-based and alignment-based tools..	56
Figure 2-2. Optimization of Mash to compute kmer-based nucleotide distance.....	59
Figure 2-3. Phages exhibit two evolutionary modes.....	61
Figure 2-4. Evaluation of genomic relationships between different types of viruses.....	63
Figure 2-5. Evaluation of genomic relationships between actinobacteriophages.....	66
Figure 2-6. Phages and phage clusters exhibit unique evolutionary trajectories.....	69
Figure 2-7. Evolutionary modes differ by phage lifestyle.....	73
Figure 2-8. Evolutionary modes are not correlated with phage tail morphotype.....	74
Figure 2-9. Evolutionary modes reflect different degrees of gene conservation.....	76
Figure 2-10. Cluster A phages exhibit two evolutionary modes.....	78
Figure 2-11. HGCF evolutionary mode is distinct in several genomic aspects.....	80
Figure 2-12. Evolutionary modes correlate with different rates of horizontal gene transfer.....	82
Figure 2-13. Evaluation of genome characteristics between evolutionary modes.....	84
Figure 2-14. Host phyla exhibit diversity in evolutionary modes.....	86
Figure 2-15. Many well-studied temperate phages are associated with HGCF.....	88
Figure 2-16. Measuring degrees of phage genetic isolation using MaxGCDGap.....	91
Figure 3-1. Bifidoprohage genomic comparison and characterization.....	114
Figure 3-2. Evolutionary relationships of bifidoprohages to other actinobacteriophages.....	116

Figure 3-3. Genomic relationships of bifidophages and their hosts.	119
Figure 3-4. Impact of mitomycin C on bifidobacterial growth.....	121
Figure 3-5. Mitomycin C induces prophage excision and circularization.	123
Figure 3-6. Mitomycin C increases sequencing coverage of <i>dnaJ₂</i> -integrated prophages.	125
Figure 3-7. Mitomycin C increases <i>dnaJ₂</i> -integrated prophage copy number.	126
Figure 3-8. Flow cytometry calibration and gating strategy.....	128
Figure 3-9. Mitomycin C induced changes in supernatant composition.	130
Figure 3-10. Complete phage particles are present in mitomycin C induced samples.	132
Figure 3-11. <i>dnaJ₂</i> -integrated prophages contain the Rin shufflon.	134
Figure 3-12. Characterization of the Rin shufflon.	136
Figure 3-13. Induced Bb423phi1 virion genome harbors multiple Rin shufflon variants.....	139
Figure 3-14. Uninduced Bb423phi1 prophage genome harbors multiple Rin shufflon variants.....	140
Figure 3-15. tRNA ^{Met} -integrated prophages contain a phase variation system.	142
Figure 3-16. <i>dnaJ₂</i> -integrated bifidophages attachment site analysis.....	144
Figure 3-17. <i>dnaJ₂</i> -integrated prophages exhibit high gene content flux.....	145
Figure 4-1. RedRock contains a <i>parABS</i> partitioning locus.	161
Figure 4-2. Characterization of predicted <i>parABS</i> systems.	163
Figure 4-3. Phylogenetic comparison of NTPase and CBP proteins.....	165
Figure 4-4. Subcluster A9 phages contain different <i>parABS</i> loci.....	167
Figure 4-5. <i>parABS</i> systems are expressed during lysogeny.	169
Figure 4-6. RedRock prophage exhibits increased copy number.	171
Figure 4-7. RedRock ParB exhibits <i>in vitro</i> binding affinity for <i>parS-L</i> and <i>parS-R</i>	173
Figure 4-8. RedRock <i>parABS</i> promotes plasmid-prophage incompatibility.	175

Figure 4-9. RedRock and Alma ParB exhibit distinct <i>in vitro</i> <i>parS</i> binding affinities.....	177
Figure 4-10. Partitioning systems promote prophage-prophage incompatibility.	179
Figure 4-11. ParB evolves at a different rate than ParA.	181
Figure 4-12. Cluster A inheritance strategy does not impact nucleotide composition.	183
Figure 5-1. Immunity system of phages in the L5 clade exhibits a genetic spectrum.....	209
Figure 5-2. Genomic relationship of D29 to other Cluster A relatives.....	211
Figure 5-3. Characterization of Cluster A immunity repressors.....	212
Figure 5-4. Characterization of Cluster A stoperators.....	213
Figure 5-5. Expression patterns of Cluster A phages during lysogeny and infection.....	215
Figure 5-6. Characterization of Trixie repressor <i>in vitro</i> binding affinity.....	218
Figure 5-7. L5 exhibits a spectrum of infection phenotypes.	220
Figure 5-8. Multiple L5 clade phages exhibit a spectrum of infection phenotypes.....	221
Figure 5-9. L5 clade phages exhibit asymmetric immunity.	222
Figure 5-10. Mesoimmunity phenotypes among L5 clade phages.	224
Figure 5-11. Quantification of asymmetric and incomplete immunity phenotypes.	225
Figure 5-12. Mesoimmunity phenotypes correlate with whole genome evolution.	227
Figure 5-13. Mesoimmunity phenotypes correlate with immunity system diversity.	228
Figure 5-14. Mesoimmunity patterns are repressor-mediated.	230
Figure 5-15. Evolutionary immunity transitions may not be linear.....	231
Figure 5-16. Summary of mutations present in defense escape mutants.....	233
Figure 5-17. Defense escape mutants exhibit varying degrees of virulence.....	235
Figure 5-18. phiTM42 exhibits strong homotypic and mesotypic virulence.....	236
Figure 5-19. phiTM39 and phiTM40 exhibit different degrees of virulence.	237

Figure 5-20. Comparison of phiTM46 and phiTM47 infection profiles.	238
Figure 5-21. phiTM38 exhibits limited virulence.	239
Figure 5-22. Evolution of immunity repressor domains.	242
Figure 5-23. Characterization of L5 engineered mutants and defense escape mutants.	243
Figure 5-24. Superinfection and immunity profiles of L5 engineered mutants.	244
Figure 5-25. phiTM4 exhibits limited virulence.	245
Figure 5-26. Genome evolution of D29 and its relatives.	249
Figure 5-27. P _{left} stoperator conservation among Subcluster A6 phages.	250
Figure 5-28. P _{left} expression and stoperator conservation among Subcluster A9 phages.	252
Figure 5-29. Toxicity of the highly expressed transcript from P _{left}	254
Figure 5-30. Extended DaVinci <i>rep-73</i> construct enhances immunity.	256
Figure 5-31. phiTM46 <i>rep</i> does not confer a dominant negative phenotype.	259
Figure 5-32. Extended phiTM46 repressor construct cannot be transformed.	261
Figure 5-33. phiTM45 DEM exhibits delayed superinfection of a Bxb1 CRS.	263
Figure A-1. PhameratorDB data pipeline and management.	285
Figure A-2. Parsing of a GenBank-formatted flat file record.	288

PREFACE

My path towards graduate school and the completion of this dissertation would not have been possible without the advice, support, and encouragement of many people, and I am indebted to all who have helped me along the way.

First, I would like to thank the members of my committee. My advisor, Graham Hatfull, has cultivated a rich and creative scientific environment through which I have developed a fascination and enthusiasm for phages and their amazing diversity. Additionally, Karen Arndt, Jon Boyle, Nathan Clark, Jeffrey Lawrence, and the late Roger Hendrix have all provided invaluable critique and feedback to help shape this dissertation into its final form.

Next, I would like to thank several former advisers. Prior to graduate school, I performed research in the labs of Frank Pugh, Joseph Martens, Karen Arndt, and Robert Coyne. Those exciting experiences lit the spark for scientific inquiry that motivated me to pursue doctoral training.

During my training, it has been wonderful to be part of the friendly and collaborative scientific culture fostered within the entire department as well as specifically within the Hatfull lab. I feel I have worked not only with colleagues but with friends, and they have all enriched my experience and contributed to my success.

I would also like to thank my entire family. They have continually supported me during this endeavor, and I am ever grateful for their encouragement and perspectives. Most importantly, I am grateful for my wife, Shar. She has supplied the patience, support, love, optimism, and inspiration I needed to embark on this path. With her help, I have endured the long days in the lab, overcome the unrelenting challenges, and crossed the finish line.

Finally, this dissertation was funded in part through the National Science Foundation Graduate Research Fellowship (grant #1247842) and the Graduate Research Opportunities Worldwide travel award.

1.0 INTRODUCTION

Viruses that infect bacteria were first discovered in the early twentieth century by Frederick Twort (Twort, 1915) and Félix D’Herelle (D’Herelle, 1917). For billions of years prior to this, these bacteriophages (“bacteria eaters”) have been entangled in an evolutionary arms race with their bacterial hosts and with other phages, and they have been dramatically impacting the world around us. Not only are phages ancient (Hendrix et al., 1999), but they are more abundant than bacteria (Bergh et al., 1989), they may be the most abundant biological entities in the biosphere (Chibani-Chennoufi et al., 2004), and some may even share common evolutionary origins with archaeal and eukaryotic viruses (Koonin et al., 2015). Although bacteria have developed an array of defenses to thwart phage infection (Doron et al., 2018; Labrie et al., 2010; Stern and Sorek, 2011), phages have developed an equally daunting array of counter-defense strategies (Andersson and Banfield, 2008; Samson et al., 2013). As a result, these viruses are incredibly diverse (Breitbart et al., 2002) and they play powerful roles in natural (Rohwer and Thurber, 2009; Suttle, 2005) and industrial (Brussow, 2001; Bruttin et al., 1997a) environments.

Over the past 100 years, phages have played myriad roles in the advancement of our understanding of molecular biology and in the development of biotechnology (Keen, 2015). They were used to discover restriction enzymes (Loenen et al., 2014) and to illustrate that DNA is heritable material (Hershey and Chase, 1952), and they were the first genomes to be fully sequenced (Fiers et al., 1976; Sanger et al., 1977). They have been indirectly used as tools for drug development (Frenzel et al., 2016), they have been directly used for therapeutic purposes (Schooley et al., 2017), and they have led to the discovery of the bacterial CRISPR-Cas system

that is now widely-used for genetic engineering (Barrangou et al., 2007). Characterizing phage diversity and understanding how they evolve can provide greater insight into basic processes of molecular biology and microbial environments as well as help create new genetic tools.

1.1 BACTERIAL DIVERSITY

Phages function and evolve within the context of their bacterial hosts, so bacterial diversity and evolution impacts phage diversity and evolution. Organisms within the domain Bacteria are incredibly diverse and are grouped into 60-90 phyla (Hug et al., 2016; Youssef et al., 2015). Even at this highest taxonomic ranking, the majority of these bacterial groups have only been identified through sequencing of environmental samples and have no cultured representatives (Hug et al., 2016; Youssef et al., 2015). Bacterial evolution is very dynamic as a result of substantial amounts of horizontal gene transfer (HGT)(Koonin and Wolf, 2008). A variety of elements facilitate transfer of genetic material between strains, including plasmids, transposons, gene transfer agents, and phages themselves (Koonin and Wolf, 2008). As a result, different types of genes exhibit different degrees of conservation and HGT (Achtman and Wagner, 2008; Konstantinidis et al., 2006; Konstantinidis and Tiedje, 2005; Wolf et al., 2016), and the lines between bacterial isolates, strains, and species are obscured. Bacteria thus are diverse and form a large genetic network (Koonin and Wolf, 2008; Varghese et al., 2015). Phages develop specificity for certain hosts within this genetic network, but they are capable of evolving new host specificities such that their “host range” is dynamic (Jacobs-Sera et al., 2012).

1.2 BACTERIAL TRANSCRIPTION

During infection, phages interact with many host-encoded factors. Depending on their stage of growth, phages develop strategies to either exploit or circumvent these factors, such as the bacterial transcription machinery. The bacterial RNA polymerase complex is responsible for transcription and consists of five factors (Haugen et al., 2008). Two subunits, β and β' , form the active site to read the DNA template and polymerize the RNA transcript, ω ensures proper folding of β and β' , and two copies of α interact with β , β' , and DNA (Browning and Busby, 2004). The complex is recruited to promoters by a sigma factor, which is a multi-domain protein that recognizes and binds DNA sequence elements. The recruited polymerase complex is stabilized at the promoter by the sigma factor, and the resulting holoenzyme is activated for transcription initiation (Browning and Busby, 2004). Sigma factors exhibit sequence specificity, and bacteria use a variety of sigma factors to control global expression programs. The primary sigma factor in *Escherichia coli*, σ^{70} , recognizes short sequences (“promoter elements”) approximately -10 bp and -35 bp upstream of the transcription start point (Browning and Busby, 2004). The assembly of the holoenzyme and initiation of transcription can be impacted by accessory factors. Through direct interaction with the polymerase complex, transcriptional activators can facilitate initiation and transcriptional repressors can prevent initiation (Browning and Busby, 2004). For example, one of the most highly characterized DNA-binding transcription factors, λ CI, binds to specific sequences (“operators”) positioned in proximity to promoter elements to sterically occlude the initiation complex from the promoter (Lewis, 2011). After initiation, the RNA polymerase complex dissociates from σ and translocates across the DNA (Santangelo and Artsimovitch, 2011).

Transcription is terminated when the elongation complex becomes displaced from the DNA template, which primarily occurs through either a rho-independent or rho-dependent mechanism (Ciampi, 2006; Peters et al., 2011). Rho-independent termination relies on the formation of RNA secondary structure (Ciampi, 2006). During elongation, sequence elements in the RNA transcript cause the polymerase complex to pause. In proximity to this pause site, a hairpin structure forms in the RNA, and extends towards the paused elongation complex. The complex is destabilized, leading to disassociation from the DNA template and termination of transcription (Peters et al., 2011). Many rho-independent terminators can be bioinformatically predicted, since they are based on the intrinsic sequence (Peters et al., 2011). In contrast, rho-dependent termination requires the factor, rho. During elongation, rho binds to RNA and translocates towards the elongation complex. Similar to rho-independent mechanisms, the elongation complex pauses due to sequence-specific elements, but during this pause rho reaches the elongation complex and destabilizes it through direct interactions, causing it to dissociate from the DNA and terminate transcription (Ciampi, 2006; Peters et al., 2011). Unlike rho-independent terminators, rho-dependent terminators are difficult to predict (Peters et al., 2011).

During infection, phages utilize and exploit host transcription machinery. However, detailed mechanistic understanding of how transcription is initiated and terminated has been gained predominantly within *E. coli* systems (Unniraman et al., 2002). As a result, transcriptional processes in other bacterial organisms may differ. For example, although *E. coli* uses 7 different sigma factors (Browning and Busby, 2004), *Mycobacterium smegmatis* encodes over 20 sigma factors (Waagmeester et al., 2005). Mycobacterial promoter elements exhibit substantial sequence variation from *E. coli*. Although mycobacterial and *E. coli* -10 elements are very similar, the mycobacterial -35 elements are less similar to *E. coli* -35 elements and are not highly

conserved (Bashyam et al., 1996; Newton-Foot and Gey van Pittius, 2013). In addition, the optimal spacing between the -10 and -35 elements in mycobacterial promoters may be wider than in *E. coli* promoters (Agarwal and Tyagi, 2006; Newton-Foot and Gey van Pittius, 2013). The diverse promoter elements suggest that mycobacteria are better able to utilize a wider variety of promoters. In general, although *E. coli* promoters are active in mycobacteria, mycobacterial promoters are not always active in *E. coli* (Agarwal and Tyagi, 2006; Newton-Foot and Gey van Pittius, 2013; Unniraman et al., 2002). Similar patterns are also observed in *Streptomyces* promoter elements (Newton-Foot and Gey van Pittius, 2013; Strohl, 1992). Lastly, mycobacteria do not contain many bioinformatically identified hairpins, suggesting rho-independent termination mechanisms are not as heavily used (Peters et al., 2011). The differences between mycobacterial and *E. coli* promoters highlight how phages that infect each host may evolve specific mechanisms to leverage each host's transcriptional initiation and termination machinery.

1.3 PHAGE EVOLUTION AND MOSAICISM

Similar to bacteria, phages exhibit substantial levels of horizontal gene transfer, but at greater levels (Hatfull and Hendrix, 2011). HGT between phages was observed through DNA heteroduplex experiments between enterobacteria phage λ and its genetic relatives (lambdoid phages), revealing that segments of lambdoid phage genomes exhibit high sequence similarity interspersed with segments of no similarity (Campbell, 1994; Casjens et al., 1992; Hershey, 1971; Highton et al., 1990). This has subsequently been confirmed by genome sequencing (Juhala et al., 2000; Ravin, 2015). Since then, this mosaic pattern of abrupt changes in sequence homology has been reported not only for other groups of enterobacteria phages (Dobbins et al.,

2004), but for phages of diverse hosts, including *Staphylococcus* (including phage Twort, originally reported in 1915)(Kwan et al., 2005), *Lactococcus* (Proux et al., 2002), *Salmonella* (Moreno Switt et al., 2013), and *Mycobacterium* (Hatfull, 2010; Pedulla et al., 2003).

The mosaic patterns arise from recombinational events between phages, prophages, and bacterial genomes such that homologous or non-homologous genes, or groups of genes, are horizontally transferred (Lawrence et al., 2002). Many mechanisms may drive HGT, but they are not well understood. Homologous recombination may occur at regions of sequence similarity as seen in other types of organisms (Lawrence et al., 2002), via host-derived or phage-derived recombinases, such as *Mycobacterium* phage Che9c gp60 and gp61 (van Kessel and Hatfull, 2008). Homeologous recombination may occur at segments that are similar, but not identical, as reported for the λ Red recombination system (Martinson et al., 2008). Additionally, it has been proposed that legitimate recombination may also occur at very short regions of homology, or illegitimate recombination may occur when no sequence similarity is present (Hatfull, 2010). The illegitimate processes may be driven by host-encoded non-homologous end joining (NHEJ) DNA repair machinery, as reported for *Mycobacterium* phages Omega and Corndog (Pitcher et al., 2006), which has also been implicated in bacterial horizontal gene transfer (Bertrand et al., 2019).

As a result of these diverse recombinational mechanisms, HGT may occur between any phages, implying that phages have access to a common pool of millions (Ignacio-Espinoza et al., 2013) or billions (Rohwer, 2003) of distinct genes, many of which currently have no known function (Edwards and Rohwer, 2005; Hatfull, 2018; Kwan et al., 2006). In this scenario, phages evolve through the genetic interactions they experience within their host range in combination with a gradual change in the host range itself (Hendrix et al., 1999; Jacobs-Sera et al., 2012). For

instance, sequence analyses suggest that phage Patience has recently shifted its host range to be able to infect *M. smegmatis* (Pope et al., 2014) and phages BP-4795 and cdtI have recently shifted their host ranges to be able to infect *E. coli* (Chithambaram et al., 2014b). By changing their host range, these phages may now have access to a different portion of the phage gene pool.

One consequence of phage mosaicism is the difficulty in which phages are directly compared. Phages that utilize different types of genetic material (dsDNA, ssDNA, etc.) do not appear to genetically interact through recombination (Lawrence et al., 2002). Beyond these broad distinctions though, there is no single DNA sequence (Rohwer and Edwards, 2002) or gene module (Lima-Mendez et al., 2008) that is shared between all phages. Discrete evolutionary lineages of phages infecting cyanobacterial hosts (Deng et al., 2014; Gregory et al., 2016), such as *Synechococcus* and *Prochlorococcus* genera, have been reported, but broad strategies to evaluate the evolutionary origin of all phages have been obfuscated due to mosaicism.

Several strategies have been developed to group phages based on their evolutionary relationships. The International Committee on Taxonomy of Viruses (ICTV) has developed a Linnaean system that heavily relies on phage morphology (Maniloff and Ackermann, 1998). Alternatively, phages have been grouped by their structural genes (Proux et al., 2002) or by “signatures” genes of variable functions (not strictly related to particle structure)(Rohwer and Edwards, 2002). Phages have been grouped using a reticulate strategy in which they may belong to multiple groups at the same time (Lawrence et al., 2002; Lima-Mendez et al., 2008). Phages infecting hosts in the phylum Actinobacteria have been grouped using a combination of factors, including nucleotide sequence similarity, gene content, and genome architecture (see below)(Hatfull et al., 2010). These varied approaches highlight the ongoing challenges of directly comparing, grouping, and evaluating phage diversity.

1.4 TEMPERATE PHAGES

One important factor that may impact how phages evolve is their lifestyle. After infection, many phages are obligately lytic (also referred to as lytic phages). They immediately enter the lytic cycle, where they replicate, produce more virion particles, and lyse the cell (Figure 1-1). However, many phages are temperate and can choose between lytic growth and a separate life cycle, lysogeny, as illustrated through the discovery and isolation of phage λ in 1951 from *E. coli* strain K-12 (Lederberg, 1951; Lederberg and Lederberg, 1953). As a lysogen, the *E. coli* host strain and “provirus” λ exhibit a “symbiotic” relationship until lytic growth of λ is induced with ultraviolet irradiation (Figure 1-1)(Lederberg and Lederberg, 1953). As a temperate phage, λ quickly became an invaluable model system for many researchers, and examination of how λ maintains lysogeny led to the discovery of basic molecular biological processes such as site-specific DNA recombination and gene regulation (Gottesman and Weisberg, 2004; Lewis, 2011; Ptashne, 1992). Phage λ was not only the first dsDNA genome to be sequenced (Sanger et al., 1982), but it also played a role in the development of the “shotgun sequencing” technique itself (Shendure et al., 2017).

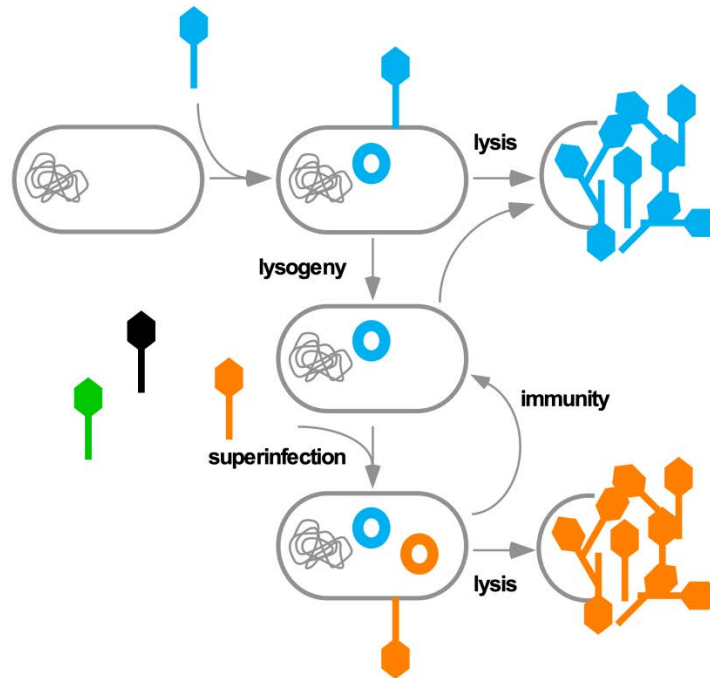


Figure 1-1. Diagram of temperate phage lifecycles.

After infection, a temperate phage (blue) may undergo lytic growth, in which new virion particles are produced and the cell is lysed. Alternatively, the temperate phage may enter a latent, non-productive lysogenic cycle in which the prophage genome is integrated into the host genome or remains as an extrachromosomal replicon similar to a plasmid. During lysogeny, the host is susceptible to a second round of infection by other phages (green, black, orange). The superinfecting phage (orange) may undergo lytic growth, or the prophage may prevent lytic growth, conferring immunity to the host against superinfection.

1.4.1 Prevalence of lysogeny

In addition to λ , many phages are temperate and exhibit viral latency. Nearly half of all isolated phages infecting *Mycobacterium* hosts are predicted to be temperate (Hatfull, 2010). Sequencing of bacterial genomes reveals that lysogeny is common (Bobay et al., 2013; Casjens, 2003), with nearly half of bacterial genomes carrying one or more prophages (Lawrence et al., 2001; Touchon et al., 2016). Sometimes up to 10-20% of bacterial genomes are derived from prophages (Brussow and Hendrix, 2002; Casjens, 2003). Even within diverse environments, such

as marine coral reefs (Knowles et al., 2016) and the murine gut microbiome (Kim and Bae, 2018), lysogens appear to be common.

Remaining within the host as a latent virus has evolutionary risks. The discovery of λ was accompanied with the realization that prophages may be susceptible to mutations that ameliorate the “pathogenic effects of the virus” such that the “virus remains trapped within the host that it never lyses”(Lederberg and Lederberg, 1953). With the increased sequencing of bacterial genomes, this has proven to be true. Analysis of over 200 prophage-like elements in a variety of bacterial strains revealed that over 95% of prophage-like elements may represent cryptic prophages, such as the CP4-44 and Rac in *E. coli* (Casjens, 2003). The *E. coli* O157:H7 strain contains 17 defective, “cryptic” prophages (Asadulghani et al., 2009). *Streptococcus pyogenes* carries 8 prophages, five of which have recognizable, substantial deletions, and only one of which is inducible (Brussow and Hendrix, 2002). Cryptic prophages present in enterobacterial strains are under purifying selection (Bobay et al., 2014), and a strain of *E. coli* K-12 carrying 9 cryptic prophages exhibits improved responses to environmental stresses (Wang et al., 2010b), indicating that trapped prophages may provide adaptive benefits to the host. Thus, the ability for temperate phages to remain as latent viruses within the host may provide unique benefits and challenges that obligately lytic phages do not encounter, and phages in these two broad categories may exhibit different patterns of diversity and evolution.

1.4.2 The genetic switch

In order to leverage the lysogenic lifecycle, temperate phages require specific types of genes and regulatory mechanisms. Unlike phages that are obligately lytic, temperate phages rely on a genetic switch that controls phage growth at three different points during its two lifecycles:

to decide whether to enter the lytic cycle or the lysogenic cycle after infection (the lysis-lysogeny decision), to maintain lysogeny as a prophage, and to re-enter the lytic cycle from a lysogenic state (the lysogeny-lysis decision)(Figure 1-1)(Oppenheim et al., 2005). Lytic growth requires controlled expression of a cascading series of transcriptional events, as observed in phages such as enterobacteria phage λ (Casjens and Hendrix, 2015), *Streptomyces* phage phiC31 (Ingham and Smith, 1992; Wilson et al., 1995), *Lactococcus* phage TP901-1 (Madsen and Hammer, 1998), and *Mycobacterium* phage L5 (Hatfull and Sarkis, 1993; Nesbit et al., 1995). The genetic switch controls initiation of this cascade at each stage using a variety of factors that form distinct, but inter-related, genetic circuits.

During infection, the genetic switch leads to lytic growth or lysogeny depending on which genetic elements are expressed. The genetic switch in enterobacteria phage λ , examined in detail over the course of several decades, harbors the most highly-characterized genetic switch (Casjens and Hendrix, 2015; Ptashne, 1992). In λ , genes required for lytic growth are expressed in several stages designated “immediate early”, “early”, and “late” (Casjens and Hendrix, 2015). The λ genome contains multiple rho-dependent and rho-independent terminators that prevent expression of genes required for lytic growth (Campbell, 1994; Casjens and Hendrix, 2015; Juhala et al., 2000). Expression of late genes is prevented due to transcriptional terminators, but these are overcome by the anti-terminator Q, which is expressed from the early gene, *Q*. Expression of early genes is prevented due to transcriptional terminators as well, but these are overcome by the anti-terminator N, which is expressed from the immediate early gene, *N*. The anti-terminators interact with the RNA polymerase elongation complex in different ways to enable it to overcome different types of terminators (Santangelo and Artsimovitch, 2011). Expression of immediate early genes is prevented by the transcriptional regulator CI. Expression

of *cI* occurs immediately after infection by the transcriptional activator CII (which is stabilized by yet another factor, CIII), but it can be repressed by Cro, which is expressed as an immediate early gene. Competing activities of Cro and CII determine whether *cI* is repressed, leading to lytic growth, or expressed, leading to lysogeny (Oppenheim et al., 2005; Ptashne, 1992).

During lysogeny, expression of *cI* leads to repression of immediate early genes *cro* and *N* and maintenance of the lysogenic state (see below for more details on CI activity during lysogeny). The prophage can re-enter the lytic cycle when this repression is ablated (Figure 1-1). In lambdoid phages, one way this can be initiated is by the *E. coli* SOS response (Kreuzer, 2013). Under normal growth conditions, several *E. coli* genes associated with DNA damage repair (SOS genes) are repressed by the DNA-binding transcription factor, LexA. In the presence of DNA damage, the co-protease, RecA, is activated and causes auto-catalytic cleavage of LexA, which de-represses SOS genes and promotes DNA repair. Similar to LexA, CI can also be deactivated by interactions with RecA, leading to expression of immediate early genes and lytic growth (Oppenheim et al., 2005; Ptashne, 1992). Many prophages can be artificially induced using DNA-damaging agents, such as ultraviolet irradiation or mitomycin C, that trigger the SOS response (Casjens and Hendrix, 2015; Lugli et al., 2016b; Oliveira et al., 2017).

1.4.3 Evolution of temperate phages

The evolutionary selective forces for phages to be able to lysogenize the host are not well understood (Chibani-Chennoufi et al., 2004), but phages have different impacts on their environment depending on their lifestyle. Lytic growth may heavily impact marine nutrient cycling (Suttle, 2005), so there may be selection for obligately lytic phages during higher densities of bacterial hosts. Similarly, obligately lytic phages are notorious for their destructive

impact on dairy manufacturing processes (Brussow, 2001). In contrast, there may also be selection for lysogeny at high bacterial densities (Knowles et al., 2016), since many temperate phages carry phage defense systems useful to the host (Bondy-Denomy et al., 2016; Montgomery et al., 2019). Spontaneous induction and release of prophages may improve the host's fitness by targeting nearby non-lysogenic competitors (Bossi et al., 2003).

As a result, phage lifestyle may play a substantial role in how phages evolve. Temperate phages may reside within host cells for longer periods of time than lytic phages (Chopin et al., 2001). Lytic phages appear less genetically diverse than temperate phages, based on analyses of phages infecting *Listeria* (Denes et al., 2014) and *Lactococcus* hosts (Chopin et al., 2001), as well as of larger collections of phages (Lima-Mendez et al., 2008). Genetic mosaicism is observed among lytic and temperate phages, but mosaicism is variable. Temperate phages may exhibit increased mosaicism due to their lifestyle (Lawrence et al., 2002), and mosaicism among lytic phages may be driven by genetic interactions with temperate phages (Dobbins et al., 2004). Lytic phages may utilize replication machinery that is less dependent on host factors (Chen and Lu, 2002). Temperate phages may depend more on genetic factors of the host, such as the host integration site and host genes required for integration. Additionally, their host range appears more restricted than lytic phages (Popa et al., 2017), they exhibit increased codon adaptation to the host compared to lytic phages (Chithambaram et al., 2014a; Lucks et al., 2008), and they have enriched sequence motifs related to host factors (Bobay et al., 2013). Overall, the genetic differences between temperate and lytic phages suggest they exhibit different patterns (modes) of evolution.

1.5 IMMUNITY SYSTEMS

In order to understand how temperate phages impact their environment and bacterial hosts, it is important to understand how they control lytic growth during lysogeny. The genetic switch controls three different stages during two lifecycles using many genetic factors. However, not all factors are needed for each stage. For example, in the λ system, CII and CIII are required to establish, but not maintain, lysogeny (Oppenheim et al., 2005).

The factors forming the genetic circuit that the phage uses to control and maintain lysogeny also directly impact how other phages are able to superinfect the lysogenic host. As a result, this part of the genetic switch constitutes the immunity system (Degnan et al., 2007; Donnelly-Wu et al., 1993; Heinrich et al., 1995; Kameyama et al., 1999; Karlsson et al., 2006). The immunity system is critical for the temperate phage to control lytic gene expression within its own genome as well as in other phage genomes. Several different types of immunity systems have been described, and some are more highly characterized than others. They exhibit many similarities, and the use of DNA-binding regulators to repress the lytic transcriptional cascade is a common, core component. However, these systems vary in complexity, and there are many differences between them.

1.5.1 Enterobacteria phage λ

Enterobacteria phage λ harbors the most highly characterized immunity system (Ptashne, 1992). This system is comprised of a single genetic locus and primarily involves expression of the two DNA-binding transcriptional regulators, *cI* and *cro*. These two factors are expressed from divergently-oriented genes, and they have similar, but non-identical, binding affinities for

two tripartite operators that consist of three 17 bp, nearly symmetric, sites (Campbell, 1994; Ptashne, 1992). One operator site, O_R , is positioned within the intergenic region between *cI* and *cro*, and the other, O_L , is ~ 2 kb downstream of *cI* between *rexB* and *N* (Casjens and Hendrix, 2015; Juhala et al., 2000).

The molecular interactions that enable CI to control transcription are complex (Little, 2010; Ptashne, 1992). Substantial work has been done to understand how a CI monomer interacts with other CI monomers to bind DNA, including a series of crystallographic studies that highlight how this factor binds as an octamer in an alternative pairwise cooperative manner (Bell et al., 2000; Bell and Lewis, 2001; Pabo and Lewis, 1982; Stayrook et al., 2008). CI exhibits an N-terminal domain that contains a helix-turn-helix DNA-binding domain, now also identified in many other transcriptional regulators (Donnelly-Wu et al., 1993; Wood et al., 1990). The N-terminus is responsible for recognizing and binding to an operator half site. The C-terminal domain enables CI to form dimers, tetramers, and octamers, which is critical to its regulatory functions. The tripartite operator O_R consists of three pairs of half sites overlapping the divergent *cI* and *cro* promoters. CI binds to two half sites as a dimer, and binds to two adjacent full sites as a tetramer through an alternative pairwise manner that prevents binding to all three full sites at the same time (Ptashne, 1992). Tetrameric CI blocks transcription initiation of the *cro* promoter and interacts with σ^{70} of the RNA polymerase holoenzyme to activate *cI* transcription initiation (Lee et al., 2012; Little, 2010). A CI tetramer also binds O_L downstream of the *cI* operon, and forms an octamer with a CI tetramer at O_R through DNA looping that enhances transcription of *cI* and prevents transcription of *N* (Ptashne, 1992).

CI binding can be disrupted by the *E. coli* SOS response (Kreuzer, 2013; Oppenheim et al., 2005). Activated RecA promotes auto-catalytic cleavage of CI between the N-terminal and

C-terminal domains in the same way that it cleaves LexA. Cleavage prevents CI from properly dimerizing and binding to operators, resulting in transcription initiation of *cro*. Cro binds as a dimer to operators, but instead blocks transcription initiation of *cI* (Ptashne, 1992). As a result, Cro expression prevents expression of *cI*, which de-represses *N* and promotes the lytic transcriptional cascade, leading to induction of λ .

1.5.2 Enterobacteria phage P2

Enterobacteria phage P2 and genetic relatives contain an immunity locus structured similarly to λ (Karlsson et al., 2006). The P2 immunity repressor, C, is located immediately upstream of the integrase, and it binds to non-palindromic tandem 8 bp half sites as a dimer (Karlsson et al., 2006; Massad et al., 2010). Genetic diversity of C and binding sites among P2-related phages correlate with repressor specificity and immunity groups (Karlsson et al., 2006).

1.5.3 *Bacillus* phage Φ 105

The immunity system of *Bacillus subtilis* phage Φ 105 has been partially characterized (Van Kaer et al., 1987). Similar to λ , *immF* contains divergently-oriented promoters. The immunity repressor, *c_{Φ105}*, is expressed from the *immF* locus. The other promoter is associated with genes required for lytic growth. Similar to λ CI, *C_{Φ105}* is a negative and positive regulator of transcription (Van Kaer et al., 1987) and contains a helix-turn-helix DNA-binding domain at the N-terminus (Van Kaer et al., 1988). The active form is a tetramer (Van Kaer et al., 1989), and unlike CI, it recognizes asymmetric, 14 bp operators.

1.5.4 *Mycobacterium* phage BPs

Integration-dependent immunity systems, such as in *Mycobacterium* phage BPs, represent a slightly less complex system than in λ (Broussard et al., 2013). The immunity locus is structured similarly to lambdoid immunity loci, in which there are two transcriptional regulators, *rep* and *cro*, expressed in opposite orientations. The immunity repressor, Rep, contains a C-terminal *ssrA*-like degradation tag (Broussard et al., 2013). During lytic growth, Rep¹³⁶ is expressed, but the tag targets Rep¹³⁶ for degradation, preventing BPs from maintaining lysogeny. During infection, the integrase is expressed, and the *attP* site is positioned within the repressor gene. During integration, the virion genome is cut within *rep*, removing the C-terminal degradation tag from the rest of the repressor. In the prophage state, a truncated isoform, Rep¹⁰³, is stably expressed since it no longer contains the degradation tag. Rep¹⁰³ binds to six operator sites, each comprised of two 12 bp half sites, located within promoter regions throughout the genome (Villanueva et al., 2015).

1.5.5 Miscellaneous λ -related systems

Immunity loci structured similarly to λ , with two divergently-oriented transcriptional regulators, have been reported in phages of diverse hosts, including *Lactococcus* phage P335 and genetic relatives (Durmaz et al., 2002), *Lactococcus* phage TP901-1 (Brussow, 2001), *Staphylococcus* phage Φ 11 (Iandolo et al., 2002), *Streptococcus* phage Φ Sfi21 (Bruttin et al., 1997b), *Lactobacillus* phages A2 and gle (Brussow, 2001), some *Listeria* phages (Kwan et al., 2005), and some *Salmonella* phages (Moreno Switt et al., 2013). Many of these systems are not well characterized, but they do exhibit some variation relative to λ .

1.5.6 Enterobacteria phage P22

Enterobacteria phage P22 contains a slightly more complex, bipartite immunity system (Susskind and Botstein, 1978). The *immC* locus contains the *c2* repressor, structured very similarly to λ *cI* in which the N-terminus is involved in binding DNA and the C-terminus is involved in dimerization and cooperativity (De Anda et al., 1983). C2 is monomeric in solution but dimerizes during binding of two symmetric half sites (De Anda et al., 1983). Divergently transcribed from *c2* is *cro* (Schicklmaier and Schmieger, 1997). C2 is the primary repressor to prevent lytic gene expression by binding two sets of operators flanking *c2* during lysogeny (Susskind and Botstein, 1978). However, it also contains a second locus, *immI*, from which an anti-repressor, *ant* is expressed (Susskind and Botstein, 1978). Ant inactivates C2 through a direct interaction (Campbell, 1994; Susskind and Botstein, 1978). At the same locus, a second repressor, *mnt*, is expressed. This DNA-binding protein binds operators to regulate *ant* expression, preventing C2 inactivation (Sauer et al., 1983). Both Mnt and C2 are required to maintain lysogeny (Susskind and Botstein, 1978).

In addition to the immunity system, P22 utilizes several mechanisms to prevent phages from superinfecting, including *sieA* (Susskind and Botstein, 1978). SieA is located at the inner cell membrane (Hofer et al., 1995) and likely blocks injection of phage DNA (Susskind and Botstein, 1978). Although these systems provide defenses against superinfection, they are not regarded as part of the immunity system since they have no role in establishing or maintaining lysogeny, and they may impact both genetically related (homotypic) and unrelated (heterotypic) phages (Susskind and Botstein, 1978).

1.5.7 Enterobacteria phage P1

Enterobacteria phage P1 (and its relative P7) contains an even more complex, tripartite, immunity system. This tripartite immunity system is a multi-layered circuit that utilizes multiple transcriptional regulators expressed from three genetic loci (*immC*, *immI*, and *immT*) (Heinrich et al., 1995). The *immC* locus is analogous to the immunity locus in lambdoid phages. At *immC*, a transcriptional regulator, *cI*, is expressed. C1 is structured differently to λ CI (Heinrich et al., 1995), and it binds to more than 15 operators within promoter regions distributed throughout the P1 genome (Lobocka et al., 2004). The operators are asymmetric, 17 bp in length, and are oriented relative to transcription, and C1 binds as a monomer (Lobocka et al., 2004). The gene *coi* (*c one inactivator*) is also expressed at this locus, and Coi non-covalently binds to C1 and prevents it from negatively regulating transcription (Heinzel et al., 1992). At the *immT* locus, the Lxc co-repressor is expressed from the *lxc* gene. Lxc enhances C1's ability to negatively regulate genes. At the *immI* locus, several genes are expressed, including an anti-repressor, *ant*, and a noncoding RNA, *c4*. Expression of C4 inhibits translation of Ant. Ant inactivates C1, potentially by directly interacting with amino acids within *sas* (*site of ant specificity*) on the C1 product. The complex interactions between factors from all three loci result in maintenance of lysogeny and superinfection immunity. The P1 genome contains a LexA binding site upstream of *coi*, and P1 may be inducible by ultraviolet irradiation (Lobocka et al., 2004). Although C1 is not homologous to λ CI, the P1 prophage stability may still be controlled through the SOS response: RecA-mediated autocatalytic cleavage of LexA leads to de-repression of *coi*, leading to inactivation of C1 and de-repression of lytic genes.

1.5.8 Enterobacteria phage N15

Similar to P1 and P7, extrachromosomal enterobacteria phage N15 contains a tripartite immunity system, consisting of three distinct genetic loci (*immA*, *immB*, and *immC*)(Ravin, 2015). *immB* is analogous to the λ *cI/cro* locus. It contains two adjacent transcriptional regulators, *cB* and *cro*, that are expressed in divergent orientations. CB performs similar functions to λ CI, binding a series of operators flanking the *cB* gene. In contrast, Cro does not appear to regulate *cB* expression as λ Cro does. The *immA* locus contains several genes, including the anti-repressor *antA* (Ravin et al., 1999). AntA interferes with CB binding through an unknown mechanism, and *antA* expression is repressed by a noncoding RNA, *cA*, analogous to the P1 *c4* noncoding RNA (Heinrich et al., 1995). The regulatory factors expressed from *immA* and *immB* during infection may determine the lysis-lysogeny switch (Ravin et al., 1999). The third locus, *immC*, encodes an anti-repressor, *antC* (Mardanov and Ravin, 2007). AntC interacts with CB *in vivo*, suggesting it directly binds to CB to interfere with regulatory activities, similar to the anti-repressor activities of Coi in P1 (Heinzel et al., 1992) and Ant in P22 (Susskind and Botstein, 1978). Unlike the lambdoid system, CB is not directly impacted by RecA-mediated autocatalytic cleavage (Ravin, 2015). Instead, expression at *immC* is controlled by the host factor, LexA, which undergoes RecA-mediated autocatalytic cleavage. During the SOS response, LexA becomes inactivated, *antC* is expressed, and CB-controlled genes become subsequently de-repressed (Mardanov and Ravin, 2007). Therefore, the interacting regulatory factors expressed from *immB* and *immC* may determine the lysogeny-lysis switch during lysogeny (Mardanov and Ravin, 2007). Several other prophages have a similar tripartite system, but they are not well-characterized, including *Klebsiella* phage Φ KO2, *Yersinia* phage PY54,

Halomonas phage ΦHAP-1, and *Vibrio* phage Vp58.5 (Casjens et al., 2004; Hammerl et al., 2016; Hammerl et al., 2015; Ravin, 2015).

1.5.9 *Streptomyces* phage phiC31

Streptomyces phage phiC31 represents another departure from the λ paradigm. The immunity repressor of phiC31 is expressed in multiple isoforms that recognize 17 bp conserved inverted repeats (CIRs) with varying binding specificities (Ingham et al., 1994; Smith and Owen, 1991). During lytic growth, expression occurs from multiple lytic promoters (Ingham et al., 1993). Throughout the genome, there are 16 CIRs, predominantly in intergenic regions, some in promoters such as those active during lytic growth, and some near terminators (Smith et al., 1999). The roles of these sites are not well understood.

1.5.10 *Mycobacterium* phage L5

The immunity system of *Mycobacterium* phage L5 is also structured much differently than λ . L5 utilizes a single transcriptional repressor (Rep) that is necessary and sufficient to confer immunity against homotypic superinfection (Donnelly-Wu et al., 1993). Rep binds to 13 bp asymmetric sequences as a monomer (Bandhu et al., 2010; Brown et al., 1997), although it can form dimers in solution (Ganguly et al., 2007), and cooperativity has not been reported. Rep contains a helix-turn-helix DNA-binding domain near the N-terminus (Donnelly-Wu et al., 1993), and the entire protein likely contains two domains similar to λ CI (Ganguly et al., 2007), suggesting that L5 prophage may be induced through auto-catalytic cleavage of Rep, mediated by RecA, similar to λ . Despite these structural similarities though, and despite the fact that

mycobacteria have an identified SOS response system with LexA and RecA homologs (Agarwal and Tyagi, 2006), L5 prophage is not inducible by DNA-damaging agents (Hatfull, 2012).

The L5 genome was sequenced and characterized in 1993 (Figure 1-2)(Hatfull and Sarkis, 1993). Many genes associated with structure and assembly of phage particles are on the top strand of the left arm of the genome. Many genes associated with DNA replication are on the bottom strand of the right arm of the genome. In the middle of the genome is the integration system. Two temporal stages of L5 expression have been reported, using a thermo-inducible L5 lysogen, in which L5 Rep becomes inactivated at elevated temperatures (Hatfull and Sarkis, 1993). After induction, early expression begins within the first 10 minutes and late expression begins after 20-25 minutes (Hatfull and Sarkis, 1993). During lysogeny and during early lytic growth, *rep* is expressed from three distinct promoters (P1, P2, and P3) within the upstream intergenic region (Figure 1-2)(Nesbit et al., 1995). Within 10 minutes after induction, expression from all three promoters diminishes, and during late lytic growth, no *rep* expression is detected (Nesbit et al., 1995). During early and late lytic growth, expression of genes on the right arm of the genome is initiated at a promoter at the right end of the genome, P_{left}. There is very little or no expression initiating from P_{left} during lysogeny.

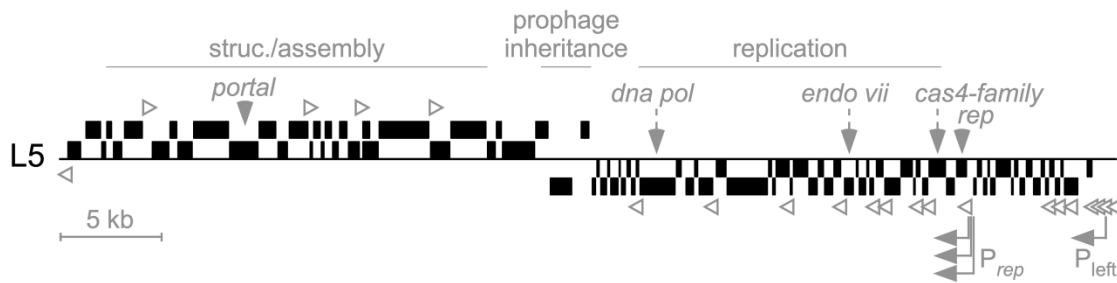


Figure 1-2. Genome map of *Mycobacterium* phage L5.

Map of the *Mycobacterium* phage L5 genome highlighting L5-specific genomic features as well as common genomic features among all Cluster A phages. Genes (black boxes) are positioned above or below the line to indicate transcriptional orientation. Many genes associated with virion structure and assembly are positioned in the left arm, many genes associated with replication are positioned in the right arm, and genes associated with prophage inheritance (integrase or partitioning) are positioned in the center. The positions of several genes highly conserved among Cluster A phages are indicated (*dna pol*: DNA polymerase, *endo vii*: Endonuclease VII, *rep*: Immunity Repressor). The early lytic promoter, P_{left} , and multiple repressor promoters, P_{rep} , are indicated by arrows, and asymmetric stopper operators are indicated by open arrowheads.

Rep is required to maintain lysogeny (Donnelly-Wu et al., 1993), and in contrast to λ it may control expression with 20-30 binding sites distributed throughout the genome (Figure 1-2) (Brown et al., 1997). Rep blocks transcription only when the asymmetric sites are oriented relative to the direction of transcription, and repression increases as the number of sites are present (Brown et al., 1997). Several binding sites are positioned in proximity to P_{left} . One overlaps the -35 promoter element and acts as a canonical operator, likely interfering with transcription initiation by RNA polymerase. The majority of sites, though, are not associated with promoters, and Rep can bind to these “stopper operators” *in vitro* to block transcription elongation (Brown et al., 1997; Rybniker et al., 2008), similar to the *Bacillus subtilis* transcriptional regulator CodY (Belitsky and Sonenshein, 2011), or the yeast transcriptional regulator Reb1 (Colin et al., 2014), both of which induce transcription termination via a roadblock mechanism.

Although the roles of many stoperators are not known, they may be involved in preventing aberrant or cryptic transcription during lysogeny. Unlike other phages such as λ , L5 does not contain many rho-independent terminators, so the stoperators may perform the analogous function to terminate transcription. For instance, many genes within the first several kilobases downstream of P_{left} are not essential (Sarkis et al., 1995). Some of these genes are cytotoxic, and they can be regulated by L5 Rep due to a stoperator site positioned between genes (Rybniker et al., 2008).

1.5.11 Superinfection immunity

The genetic circuit that the prophage uses to control growth during lysogeny is referred to as the immunity system due to a second function that it performs. During lysogeny, the host cell remains susceptible to a second round of infection (superinfection) by other phages in the environment (Figure 1-1). Under some circumstances, regulatory elements encoded in the genetic circuit and utilized by the prophage to maintain lysogeny also provide a defense mechanism to the host against other phages, such that the host is immune to superinfection. Investigation of the λ immunity system has established the paradigm for superinfection immunity.

In the λ system, superinfection immunity is repressor-mediated (Campbell, 1994). CI exhibits binding specificity for cognate operator sites to repress *cro* and *N*. If a superinfecting phage contains a similar genetic switch with similar operators, the prophage CI can recognize these operators in the superinfecting genome and prevent expression of its lytic genes as well. Enterobacteria phages HK97 and λ harbor very similar (homotypic) immunity systems, and a λ

prophage confers immunity to the host against superinfection from both phages since CI can recognize their lytic gene regulatory elements and prevent lytic growth (Juhala et al., 2000). These two phages exhibit homoimmunity since the reciprocal challenge has the same results; HK97 can prevent λ superinfection.

Although the immunity system is required for lysogeny, it presents a target for prophage defense from homoimmune prophages. As a result, there is likely to be pressure to evolve new immune specificities such that the regulatory elements within a phage's immunity system no longer interact with the elements in other systems (Campbell, 1994). Phages can escape homotypic immunity by acquiring mutations that disrupt this circuitry (Bronson and Levine, 1971; Heinrich et al., 1995; Yarmolinsky, 2004). λ requires as few as 3 point mutations within operators to superinfect a λ lysogen; the prophage-expressed CI is unable to recognize the mutant operators and prevent lytic gene expression (Campbell, 1994). These mutations render the mutant λ to be obligately lytic since its cognate CI also fails to recognize the mutant operators and is unable form stable prophages. However, many lambdoid phages have been identified that harbor evolutionarily diverged (heterotypic) derivatives of the same regulatory circuitry and are no longer subject to each other's circuitry (Campbell, 1994; Kameyama et al., 1999). Enterobacteria phages 434 and λ harbor homologous circuitry, but their CI repressors exhibit specificity for different operator sequences and are unable to block *cro* expression in the opposing phage and are thus heteroimmune (Campbell, 1994; Ptashne, 1992). The evolution of immunity systems results in phages that form immunity groups. Phages in the same group harbor homotypic immunity systems and are homoimmune. Phages within the group are heteroimmune with phages outside of the group that contain heterotypic immunity systems.

Immunity groups have been identified for many types of phages. Among lambdoid phages, nearly 10-20 different immunity groups have been identified (Campbell, 1994; Kameyama et al., 1999). Among P2-related phages, 7 immunity groups have been reported (Karlsson et al., 2006). Enterobacteria phages P1 and P7 are heteroimmune (Heinrich et al., 1995). Several L5-related *Mycobacterium* phages exhibit distinct immune specificities. L5 is grouped as a “Cluster A” actinobacteriophage (see below). Many other phages are related to L5 and are also grouped into Cluster A. These phages exhibit similar genomic architecture and are predicted to utilize homologous immunity systems (Figure 1-2). Phage Bxb1 is in Cluster A, it exhibits a similar temporal expression program as L5 (Mediavilla et al., 2000), and its immunity system has been compared to L5 (Jain and Hatfull, 2000). Similar to L5, the Bxb1 genome contains 20-30 stoperators, but the Bxb1 stoperator consensus sequence differs from the L5 stoperator consensus sequence. Bxb1 Rep and L5 Rep both exhibit stronger specificity *in vitro* for cognate sites than non-cognate sites, and these two phages are heteroimmune, representing distinct immunity groups (Jain and Hatfull, 2000). The immunity repressor and stoperator sites can be identified in nearly all Cluster A phages. It is not clear which particular stoperator and operator sites are needed to maintain lysogeny or to confer superinfection immunity, and it is not clear whether other genetic elements are involved in immunity or required to maintain lysogeny. Nevertheless, superinfection immunity phenotypes observed between a dozen Cluster A phages correlate with repressor phylogenies and stoperator consensus sequences, and they reflect multiple immunity groups (Ford et al., 1998; Jain and Hatfull, 2000; Pope et al., 2011b).

1.6 INHERITANCE SYSTEMS

In addition to encoding genetic systems that control lytic gene expression, temperate phages must also encode genetic systems to ensure their genomes are reliably replicated and propagated to host progeny as the host grows and divides. There are two primary inheritance strategies: integrating into the host chromosome or remaining as an extrachromosomal plasmid-like replicon.

1.6.1 Integration systems

Many characterized temperate phages utilize integration systems to promote prophage inheritance, such as in phage λ , which is one of the most well-characterized systems (Grindley et al., 2006; Landy, 1989). These systems rely on site-specific recombinases, integrases, as well as other factors encoded by the phage or host (such as Xis, IHF and FIS) to facilitate the integration of the phage genome into the bacterial chromosome as well as excision to re-form the virion genome (Grindley et al., 2006). Recombination specifically occurs between the *attP* site within the phage genome and the *attB* site within the host genome, and results in the phage genome integrated within the host genome as a prophage between *attL* and *attR* sites. The latent, integrated prophage is replicated and propagated along with the host genome until induction, at which point recombination occurs between the *attL* and *attR* sites, the prophage genome is excised and circularized, and lytic growth commences (Grindley et al., 2006).

Integrases are grouped into tyrosine or serine families based on their catalytic domains and molecular basis for integration (Grindley et al., 2006). The majority of temperate *Mycobacterium* phages contain tyrosine-family integrases (Hatfull, 2012). In these integration

systems, the phage *attP* site is typically located near *int*, and they tend to utilize *attB* sites located within tRNA genes, such as the well-characterized L5 integrase which integrates into tRNA^{Gly} (Hatfull, 2012; Lee et al., 1991; Pena et al., 1997). Several temperate *Mycobacterium* phages contain serine-family integrases, and these tend to integrate within coding genes, such as the well-characterized Bxb1 Int, which utilizes an *attB* within *groEL1* (Hatfull, 2012; Kim et al., 2003). The integrases identified in temperate *Mycobacterium* phages are diverse, and over a dozen different *attB* sites in *M. smegmatis* have been identified (Hatfull, 2012).

1.6.2 Partitioning systems

An alternative strategy to ensure prophage inheritance is to remain as an extrachromosomal replicon. Few extrachromosomal prophages have been reported, so this strategy appears less common. Remaining as an extrachromosomal replicon requires the phage to encode a partitioning system. Partitioning systems have been characterized in bacterial, plasmid, and phage genomes (Livny et al., 2007; Pinto et al., 2012; Wang et al., 2013). They distribute copies of replicons to progeny during cellular growth and division to ensure replicons are stably and reliably inherited. Enterobacteria phages P1 and P7, isolated in 1972 (Smith, 1972), and N15, isolated in 1964 (Ravin, 2015), use partitioning systems to remain as extrachromosomal prophages in the host cell during lysogeny similar to plasmids. During lysogeny, prophage P1 is maintained at ~ 1 copy per cell, similar to integrated prophages (Lobocka et al., 2004), and prophage N15 is maintained at ~ 3-5 copies per cell (Ravin, 2015). Characterization of these two temperate phages, as well as of enterobacteria phage P7, substantially contributed to understanding how partitioning works (Abeles et al., 1985; Hayes and Austin, 1993; Lobocka et al., 2004).

Partitioning systems consist of two genes, a nucleotide triphosphatase (NTPase) and centromere-binding protein (CBP), in an operon that is flanked on one or both sides by multiple binding sites, and the majority of systems form a self-contained module (Ebersbach and Gerdes, 2005; Gerdes et al., 2000; Schumacher, 2012). The CBP binds sites on the replicon to form a partitioning complex that interacts with the NTPase (Schumacher, 2012). The NTPase, typically the first gene in the operon (Gerdes et al., 2000), builds a scaffold to actively segregate replicons.

There are three main types of partitioning systems, based on the structure of the partitioning system and segregation strategy (Ebersbach and Gerdes, 2005; Schumacher, 2012). Type I systems contain ATPases that harbor Walker A box domains and that polymerize and “pull” replicons to opposing sides of dividing cells (Schumacher, 2012). These systems are further subdivided into Type Ia and Ib based on size and sequence of the NTPase and CBP (Schumacher, 2012) and the position of the centromeric binding sites. Type Ia NTPases, such as ParA in P1 (Lobocka et al., 2004) or SopA in N15 (Ravin, 2015), contain a helix-turn-helix DNA-binding domain, and they auto-regulate expression through transcriptional repression at the *par* promoter. Type Ia CBPs, such as ParB in P1 (Lobocka et al., 2004) or SopB in N15 (Ravin, 2015), also contain a helix-turn-helix DNA-binding domain, but they bind to centromeric sites either immediately downstream of the *par* operon, such as the *parS* locus in P1 (Lobocka et al., 2004), or to sites distributed throughout the genome, such as the *sopC*-like loci in N15 (Ravin, 2015). In contrast, Type Ib NTPases, such as ParF in plasmid TP228 or ParA in plasmid pB171, do not contain a helix-turn-helix motif. Type Ib CBPs, such as ParG in TP228 or ParB in pB171, contain a ribbon-helix-helix DNA-binding domain (Schreiter and Drennan, 2007; Schumacher, 2012). They bind to centromeric sites either upstream of the operon, such as in TP228, or at both ends of the operon, such as *parC1* and *parC2* loci in pB171 (Ebersbach and Gerdes, 2005). Type

II NTPases, such as ParM in plasmid R1, form actin-like filaments that “push” replicons to opposing sides of dividing cells. Type III NTPases, such as TubZ in plasmid pBtoxis (Larsen et al., 2007), polymerize and generate a “tram” for replicons to traverse to opposing sides of dividing cells. Type Ib, II, and III systems are auto-regulated by the CBP at sites within the *par* operon promoter (Schumacher, 2012).

1.7 IMPACT ON BACTERIAL HOST

Bacteria interact with diverse organisms in the environment, and the evolutionary pressures imposed by these interactions directly impact how temperate phages evolve, since they reside within the host. Temperate phages carry diverse genes not directly associated with prophage inheritance, lysogeny, or lytic growth, and these genes impact their bacterial hosts and their environment in diverse, complex ways.

Some genes are associated with enhancing the phage’s fitness at the expense of the host. This is observed in the extrachromosomal phage P1’s toxin-antitoxin system encoded by *phd* and *doc* genes (Jensen and Gerdes, 1995). In the absence of Phd, Doc is toxic to the host. Both genes are expressed in an operon but Doc is more stable than Phd. If phage P1 is not successfully inherited, Doc is activated and causes cell death. A potentially new class of toxin, MuF, is present in many temperate phages and may also be involved in conferring cell death in the absence of the phage (Jamet et al., 2017).

Many temperate phages also carry genes that provide the host with defense against heterotypic phages. In the host *Pseudomonas aeruginosa*, prophages JBD23 and JBD30 prevent replication of superinfecting phage JBD88a, and prophage JBD26 modifies the host’s cell

surface to block adsorption of phage JBD24 (Bondy-Denomy et al., 2016). In the host *M. smegmatis*, prophage Fruitloop expresses gp52, which interacts with the host protein Wag31 to prevent superinfection of phages Hedgerow and Rosebush (Ko and Hatfull, 2018). In the host *Gordonia terrae*, prophage CarolAnn expresses gp43 and gp44 to defend against superinfection of phage Kita (Montgomery et al., 2019).

Many temperate phages also express genes associated with bacterial pathogenicity (Brussow et al., 2004). For instance, cholera, which has afflicted the human population for hundreds of years and which continues to infect millions of people each year, is caused by the bacterial pathogen, *Vibrio cholera* (Lippi et al., 2016). Pathogenicity is due to cholera toxin, produced by two genes, *ctxA* and *ctxB*, encoded in the prophage CTX Φ that has integrated into the host genome (Brussow et al., 2004; Waldor and Mekalanos, 1996). Similarly, botulism is a rare disease resulting from neurotoxins produced by *Clostridium botulinum* (Sobel, 2005). There are seven types of neurotoxins, and prophages (such as c-st) encode the C1 type toxin gene as well as other regulators and co-factors required for toxin expression (Brussow et al., 2004; Sakaguchi et al., 2005; Sobel, 2005). Shiga toxin-producing *E. coli*, which lead to diarrheal disease in 3 million incidences per year (Bryan et al., 2015), are a variety of *E. coli* strains carrying prophages that encode shiga toxin (*stx*) genes (Brussow et al., 2004). *E. coli* strain O157:H7 contains two prophages, VT1-Sa and VT2-Sa, and *stx1* and *stx2* genes are expressed during prophage induction (Matsushiro et al., 1999). Temperate phages are also involved in diphtheria, *Salmonella* food poisoning, and *Staphylococcus aureus* infections (Brussow et al., 2004).

In addition to encoding toxins, temperate phages may play other roles in their hosts. They may facilitate horizontal gene transfer of drug resistance genes, such as several cryptic prophages

in *E. coli* (Asadulghani et al., 2009; Wang and Wood, 2016) or prophage $\Phi 11$ in *S. aureus* (Haaber et al., 2016). Integration of phage $\Phi 11$ and $\Phi 80\alpha$ into *S. aureus* increases biofilm formation (Fernandez et al., 2018). Integration of Bxb1 into the *groEL1* gene of *M. smegmatis* leads to inactivation of *groEL1* and prevention of biofilm formation (Ojha et al., 2005). Additionally, entire prophages may be controlled by the host as large genetic switches, as observed for prophage A118 in *Listeria monocytogenes* during infection of mammalian cells (Feiner et al., 2015).

1.8 BIOTECHNOLOGICAL APPLICATIONS

Temperate phages have enabled the development of a variety of genetic tools for a variety of goals. Enterobacteria phage N15 has been used to develop several linear plasmid vectors that provide some advantages over circular plasmid vectors (Ravin, 2015). The *cre-lox* site-specific recombination system present in enterobacteria phage P1 has been widely used to study yeast genetics (Sauer, 1987). Enterobacteria phage λ 's immunity repressor, CI, has been used to develop yeast two-hybrid systems to examine protein-protein interactions (Serebriiskii et al., 1999). The serine integrases of *Streptomyces* phage phiC31 and *Mycobacterium* phage Bxb1 have been used to develop tools to genetically manipulate eukaryotic model organisms, such as *Drosophila* (Bischof et al., 2007; Huang et al., 2011). The emerging field of synthetic biology has utilized several integrases, including the well-characterized serine integrases in *Streptomyces* phage phiC31 and *Mycobacterium* phage Bxb1, and the λ *cI/cro* cassette to develop artificial genetic circuits (Bonnet et al., 2012; MacDonald and Deans, 2016; Yang et al., 2014). Several actinobacterial integration vectors using serine integration systems from *Streptomyces* temperate

phages have been developed (Baltz, 2012). The isolation of thousands of phages infecting *M. smegmatis* have led to the development of a suite of site-specific integrating vectors that can be used to study *Mycobacterium tuberculosis* (Hatfull, 2014). Temperate phages themselves have been used to manipulate antibiotic resistance of their bacterial hosts as a potential tool to combat the spread of antibiotic resistant bacterial pathogens (Monteiro et al., 2018).

1.9 ACTINOBACTERIOPHAGE DIVERSITY

Phages infecting hosts in the phylum Actinobacteria (“actinobacteriophages”) represent one of the largest collections of isolated, sequenced, and well-characterized phages. Actinobacteria represents one of the largest bacterial phyla (Barka et al., 2016). Members of Actinobacteria are gram positive, their genomes have high GC% content (ranging from 50-70%), they are genetically diverse, and they are sufficiently distinct from other bacteria such that the nearest evolutionarily related phylum is not obvious (Barka et al., 2016; Ventura et al., 2007). Members of Actinobacteria are grouped into 6 classes, nearly 40 families, and 130 genera, and they inhabit a variety of aquatic and terrestrial environments (Barka et al., 2016; Ventura et al., 2007). Although many genera are aerobic (such as *Mycobacterium* and *Streptomyces*), some genera are anaerobic (such as *Bifidobacterium* and *Propionibacterium*)(Barka et al., 2016; Ventura et al., 2007). Some genera, such as *Streptomyces*, have been useful for developing antibiotics (Procopio et al., 2012) or for industrial purposes (Nakashima et al., 2005), and others have a positive impact on human health, such as *Bifidobacterium* present in the human gut (Arboleya et al., 2016). Many are pathogenic though. *M. tuberculosis* is the causative agent of tuberculosis, causing over 8 million incidences in 2012 (Organization, 2013). *Propionibacterium*

are present on human skin and are associated with acne inflammation (Kim et al., 2002). *Tropheryma* is the causative agent of Whipple's disease while *Corynebacterium diphtheria* is the causative agent of diphtheria (Barka et al., 2016).

The first actinobacteriophage to be sequenced was L5 in 1993, infecting *M. smegmatis* (Hatfull and Sarkis, 1993). Since then, thousands of actinobacteriophages have been systematically isolated and sequenced through the Phage Hunters Integrating Research and Education (PHIRE) and Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) educational programs (Hanauer et al., 2017; Hanauer et al., 2006; Jordan et al., 2014). In these programs, phages are isolated and purified, their genomes are sequenced, and their genes are carefully, manually annotated. This refined, rigorous, process coordinates the efforts of thousands of undergraduate students, educators, and researchers from over 100 institutions, and has resulted in the isolation of 15,000 phages and the sequencing of 2,800 genomes (<http://phagesdb.org>).

Currently, there are 2,900 sequenced actinobacteriophages, isolated from hosts representing 14 of the 130 total genera (such as *Mycobacterium*, *Gordonia*, *Streptomyces*, and *Arthrobacter*) in the phylum Actinobacteria. Over 1,700 infect a single strain, *M. smegmatis* mc²155, a genetically tractable, fast-growing mycobacterial model organism competent for efficient transformation (Snapper et al., 1990). The 300,000 annotated actinobacteriophage genes are grouped by sequence similarity into over 24,000 phamilies (“phams”)(of which 35% are “orphams”, containing only a single gene member) using the data analysis pipeline to maintain the Phamerator database (Appendix A). Actinobacteriophages are evaluated and grouped into clusters to reflect genomic relationships (Hatfull et al., 2010). Several metrics are utilized, including dot plot comparison, pairwise average nucleotide identity (ANI), visual analysis using

BLAST-based whole genome alignments in Phamerator, and gene content analysis using SplitsTree. Phages grouped into the same cluster exhibit dot plot sequence similarity spanning more than 50% of their genomes, average nucleotide identities above ~ 53 -60%, close association of gene content based on SplitsTree, and gene synteny using Phamerator. Phages with no genetic relatives remain as singletons. Phages within clusters are further subdivided into subclusters if there is sufficient genetic diversity, such that phages between subclusters exhibit ANI above $\sim 64\%$ and phages within subclusters can exhibit ANI greater than 99%. The complete collection of actinobacteriophages have been grouped into 120 clusters and 71 singletons. Nearly 600 phages infecting *Mycobacterium* (25-30% of all *Mycobacterium* phages) are genetically related and are grouped into Cluster A.

1.10 CURRENT CHALLENGES

Since the isolation of λ , investigations of temperate phage diversity and evolution have led to innumerable and invaluable biological insights. However, many paradigms regarding how temperate phages evolve and impact their environment, including superinfection immunity and prophage inheritance, were primarily derived from relatively small collections of phages infecting enterobacteria, many of which are still not completely sequenced (such as the lambdoid phages 434, 21, and 82). More recent studies examining immunity systems of incompletely sequenced λ -related (Degnan et al., 2007; Kameyama et al., 1999) or P2-related phages (Karlsson et al., 2006) continue to encounter the same challenges of connecting phenotype to genotype. Since phages exhibit enormous diversity and mosaicism, limited collections of phages with incompletely sequenced genomes may only provide insight into broad evolutionary patterns

that span large genetic distances (Figure 1-3A). In contrast, larger collections exhibiting a spectrum of diversity may provide better resolution to understand incremental or gradual evolutionary processes (Figure 1-3B).

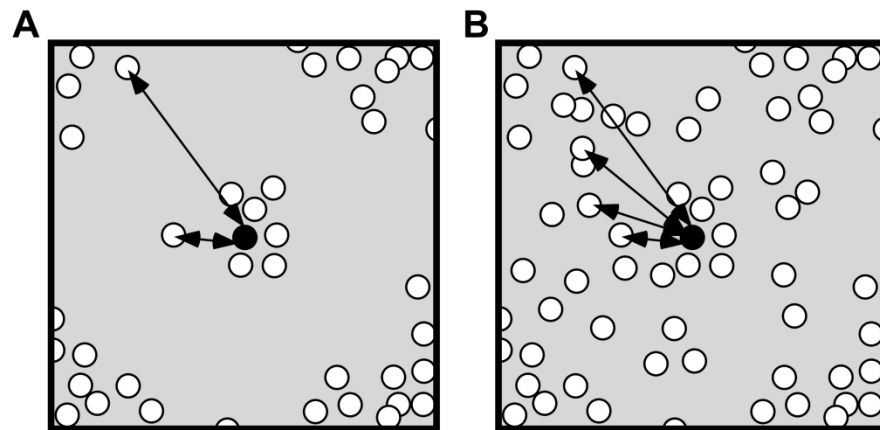


Figure 1-3. Diversity across a phage genome landscape.

Theoretical two-dimensional phage genome landscape, with (A) few or (B) many phages (circles) isolated and sequenced. (A) With few phages available, genotypic or phenotypic analyses for a specific phage (black) are limited to closely-related phages (group of circles near center, short arrow) or to very dissimilar, unrelated phages (circles at edges of landscape, long arrow). (B) With many phages available, analyses can be performed using a broader spectrum of genetic diversity (arrows of increasing length).

Fewer than 350 enterobacteria phages have been completely sequenced, and they have been systematically clustered similar to actinobacteriophages (Große and Casjens, 2014). These phages, infecting hosts spanning 18 genera, can be grouped into 38 clusters, with 18 remaining as singletons. Nearly 25% of phages in this dataset are considered lambdoid, but they represent 17 clusters or singletons, and only 3 phages are grouped in the λ cluster. P22-related phages comprise the largest lambdoid cluster, but it is only comprised of 16 phages that are further subdivided into 2 subclusters.

The current genetic diversity of sequenced actinobacteriophages far exceeds the diversity of enterobacteria phages. Although sequenced actinobacteriophages infect hosts representing only 14 genera, there are nearly 10 times more sequenced phages, and they are grouped into nearly 4 times more clusters/singletons. Phages in the largest cluster, Cluster A, harbor the most highly characterized actinobacteriophage immunity system, are nearly five times as abundant as lambdoid phages, and exhibit sufficient genetic diversity to be further subdivided into 19 subclusters.

The diversity of sequenced actinobacteriophages within and between hosts provides an opportunity to expand and refine our understanding of how temperate phages evolve at various scales, from the evolution of specific systems used for prophage inheritance or to maintain lysogeny, to the evolution of the entire genome. Instead of being limited by a small collection of phages, we now can leverage large collections of phages that reflect a spectrum of genetic diversity to investigate these evolutionary processes within greater precision (Figure 1-3B)(Pope et al., 2015).

For my dissertation, I investigate processes of temperate phage evolution at various scales, with a focus on actinobacteriophages. In Chapter 2, I compare and contrast changes in gene content (gene content flux) between temperate phages of the same or different hosts, as well as between temperate and lytic phages. In Chapter 3, I expand our understanding of temperate phages infecting *Bifidobacterium* hosts by characterizing and inducing several predicted bifidobacterial prophages. In Chapter 4, I characterize the partitioning systems in Cluster A *Mycobacterium* phages, representing the first analysis of actinobacterial temperate phages that rely on partitioning for prophage inheritance. In Chapter 5, I refine our understanding of immunity system evolution by investigating how the *Mycobacterium* Cluster A immunity system

functions and evolves new specificities. Finally, in Appendix A, I describe my contributions to maintaining and improving one of the primary SEA-PHAGES databases, PhameratorDB, used to store and evaluate actinobacteriophage genomics data. Many of the comparative genomics analyses I performed relied on this database, and I substantially contributed to its development.

2.0 INVESTIGATION OF WHOLE GENOME EVOLUTION

The data presented in this chapter relating to phage evolutionary modes were published in the journal *Nature Microbiology* (Mavrich and Hatfull, 2017). The data in this chapter relating to phage genetic isolation were published in *MBio* (Pope et al., 2017). I performed the experiments and analyses.

2.1 INTRODUCTION

Phage evolution is likely impacted by many factors. Genetic mosaicism results from frequent horizontal gene transfer (HGT) arising from legitimate [homology-based (Lawrence et al., 2002) or homeologous-based (Martinsohn et al., 2008)] and illegitimate [non-homology-based (Hatfull, 2010)] recombinational events and has been observed among phages infecting diverse host genera, including *Escherichia* (Juhala et al., 2000; Ravin, 2015), *Staphylococcus* (Kwan et al., 2005), *Lactococcus* (Proux et al., 2002), *Salmonella* (Moreno Switt et al., 2013), and *Mycobacterium* (Hatfull, 2010; Pedulla et al., 2003). Phage diversity is not homogeneous though. Among the 1,700 isolated and sequenced phages infecting *Mycobacterium*, 33% have been grouped into a single cluster (Cluster A) and 9 phages still have no close genetic relatives (<http://phagesdb.org>). Phages infecting *Arthrobacter* hosts exhibit similar degrees of genetic diversity as phages infecting *Mycobacterium* (Klyczek et al., 2017), but phages infecting *Propionibacterium* hosts exhibit very little sequence diversity (Marinelli et al., 2012). Furthermore, phages infecting hosts in the phylum Cyanobacteria, including *Synechococcus* and

Prochlorococcus genera, also exhibit mosaicism, but distinct phylogenetic lineages can be identified (Deng et al., 2014; Gregory et al., 2016), suggesting they may be subjected to different evolutionary constraints than other phages. Lastly, phage lifestyle also impacts phage evolution, with temperate phages exhibiting differences from lytic phages in gene content (Chen and Lu, 2002; Chopin et al., 2001; Denes et al., 2014; Lima-Mendez et al., 2008), codon usage (Chithambaram et al., 2014a; Lucks et al., 2008), and host-related sequence motifs (Bobay et al., 2013).

Investigating the complexities of phage evolution can help to elucidate the ways they impact their environment. HGT introduces or removes genes from vertically inherited genomes, and one strategy to evaluate patterns of phage evolution is to compare this gene content flux relative to changes in whole genome nucleotide sequence. In order to investigate general patterns of phage evolution, I developed a comprehensive, unbiased bivariate analysis using quantitative metrics of gene content and nucleotide similarity that can be used to compare any two phages. I investigated thousands of phages spanning nearly 10 host phyla. Phages exhibit two general patterns of change in gene content relative to their nucleotide sequence. These evolutionary modes (Fitch and Ayala, 1994) correlate with phage lifestyle; lytic phages are constrained to one evolutionary mode, but temperate phages exhibit both evolutionary modes. Groups of genetically related phages appear constrained to one mode or the other, and different proportions of phages exhibit the two evolutionary modes when they are grouped by host phyla.

2.2 MATERIALS AND METHODS

2.2.1 Phages used in this study

A total of 2,333 microbial viruses were used for this study [Supplementary Table 2-1, originally published in (Mavrich and Hatfull, 2017)]. All 1,941 viruses in the NCBI RefSeq database (<https://www.ncbi.nlm.nih.gov/refseq/>) listed as “microbial viruses” were downloaded on 8/11/2016 and combined with phages in a local Phamerator database *Actinobacteriophage_785*. This database contains 2,191 double-stranded DNA bacteriophages, 84 bacteriophages with alternative nucleic acid genomes (ssDNA, dsRNA, ssRNA), 23 bacteriophages with unspecified nucleic acid type, 3 archaeal viruses, 1 eukaryotic virus, and 31 viruses of unspecified origin. All 2,333 viral genomes were used to create the *Bacteriophages_2333* Phamerator database and is available online (http://phamerator.webfactional.com/databases_Hatfull). CDS features for actinobacteriophages were retrieved from the *Actinobacteriophage_785* database and CDS features for all other genomes were retrieved from their RefSeq records. Genes are grouped into phamilies (“phams”) in Phamerator using kClust (Hauser et al., 2013), forming 62,363 unique phams.

2.2.2 Collection of virus metadata

Several types of metadata were used for analysis: phage clusters, host taxonomy, viral taxonomy, and phage lifestyle (Supplementary Table 2-1). For all metadata fields, missing/incomplete data were listed as “Unspecified” and were excluded from each specific type of analysis. Host genera of the 785 actinobacteriophages were derived from PhagesDB

(<http://phagesdb.org>). For the other 1,548 genomes, the *Biopython* (Cock et al., 2009) package was used to retrieve host genus from each RefSeq record. Host data are stored in multiple fields in RefSeq records, and manual assessment of host data parsed from these fields was required to identify the appropriate host genus. For all 2,333 genomes, the *ete3* python package (Huerta-Cepas et al., 2010) was used to retrieve complete host taxonomy. For viral taxonomy, taxonomic data of the 785 actinobacteriophages were derived from PhagesDB, and for all other phages from each RefSeq record using *Biopython*. Cluster data was obtained from PhagesDB. Empirical lifestyle data for 1,067 phages was obtained from multiple sources. For the 785 actinobacteriophages in PhagesDB, lifestyle for all phages within a specific cluster was assigned if there was empirical data or strong, reasonable genomic evidence indicating they are temperate or lytic (personal communication with Welkin Pope). For phages of other host phyla, lifestyle data was compiled from two online resources, PHACTS (McNair et al., 2012) and ACLAME (Leplae et al., 2010), as well as several other previously compiled data (Chithambaram et al., 2014a; Chopin et al., 2001; Grose and Casjens, 2014; Klumpp and Loessner, 2013; Sau et al., 2005). Lifestyle data was thus curated for over 40% of the phages in this database, resulting in 452 lytic and 614 temperate phages, and predominantly from the host phyla Actinobacteria (562 phages), Firmicutes (131 phages), and Proteobacteria (362 phages).

2.2.3 Categorization of phams into general gene functions

Phams present within 785 actinobacteriophages, which have been predominantly manually and systematically curated through the SEA-PHAGES program (Jordan et al., 2014), were manually grouped into mutually exclusive, but non-exhaustive, functional categories based on the pham descriptions present in the database. Common gene functions associated with each

category are as follows: lysis (endolysin, holin, LysA, LysB, LysM, lysin); lysogeny (immunity repressor, integrase, parA, parB, excise); recombination/replication (DnaB, DnaC, DnaJ, Ftsk, helicase, ku, DNA polymerase, primase, RDF, RecA, RecB, RecE, RecT, RusA, RuvC, resolvase); structural/assembly (capsid, capsid maturation protease, capsid morphogenesis, head assembly, head-to-tail connector, head decoration, major capsid, major tail subunit, minor tail subunit, portal, terminase, structural, tail assembly chaperone, tail fiber, tape measure, scaffolding, tail, tail sheath, tail spike, baseplate); misc. (any phams with gene functions that do not clearly fit into any of the above categories); no known function (no pham description data available).

2.2.4 Prediction of phage lifestyle

The lifestyles of all phages in the dataset were predicted in order to complement the empirical lifestyle dataset (Supplementary Table 2-1). Phamerator identifies conserved domains in all genes using the NCBI conserved domain database (CDD)(Marchler-Bauer et al., 2011). All conserved domains in the dataset that contain descriptions relating to “integrase” (for integrating temperate phages) or “parA” (partitioning gene found in extrachromosomal temperate phages, see Chapter 4), or those that are associated with phams that contain manual descriptions relating to “integrase” or “parA”, were manually identified, resulting in 206 “temperate phage” domains. All phams in the dataset containing at least one temperate phage domain were identified, resulting in 149 “temperate phage” phams. All phages in the dataset containing at least one temperate phage pham were identified, resulting in 962 predicted temperate and 1,371 predicted lytic phages. The predicted lifestyle data conflicts with the empirical lifestyle data in 4% of the empirical dataset, some of which are readily identifiable as recent lytic mutant derivatives of

temperate parents such as the Cluster K phage TM4 (Pope et al., 2011a). Additionally, temperate Mu-like phages integrate into the host genome using transposases instead of integrases, so they were not identified using this strategy, accounting for several other phages on this list. Additionally, some recombinase genes may contain similar conserved domains as found in integrases, and true lytic phages containing these genes would be erroneously categorized as temperate. Conversely, true temperate phages that contain novel integration machinery would be missed by this strategy and would be erroneously categorized as lytic.

2.2.5 Calculation of whole genome gene content dissimilarity

Custom python scripts were developed to compute a gene content dissimilarity index [Supplementary Table 2-2, originally published in (Mavrich and Hatfull, 2017)]. For each pairwise comparison of genomes, the number of shared phams between the two genomes was computed, and this was divided by the total number of phams present in each genome. The two proportions were averaged and converted to a dissimilarity: $1 - \text{average shared pham proportion}$, ranging from 0 (all phams are identical) to 1 (no phams are identical). Gene function-specific gene content dissimilarity was computed for each pairwise comparison in the same manner, except that only the subset of phams grouped into each specific category were used.

2.2.6 Calculation of whole genome nucleotide distance

Mash software (Ondov et al., 2016) v1.1 was used to compute nucleotide distance between all genomes using custom bash and python scripts (Supplementary Table 2-2). For Mash optimization, pairwise ANI for all genomes in the training set was computed using default

settings in DNA Master (<http://cobamide2.bio.pitt.edu>), which implements an alignment-based approach similar to that previously described (Konstantinidis and Tiedje, 2005). Correlations of ANI-based distance ($1 - \text{ANI}$) and Mash-based distance were analyzed in RStudio (<https://www.rstudio.com>) to determine the optimal parameters. After optimization, the following parameters were used for to compare all genomes in the database: kmer = 15, sketch = 25,000, $p\text{-value} < 1 \times 10^{-10}$, and genome size disparity of 100%. Fewer than 2% of all comparisons are impacted by the genome size disparity parameter, and fewer than 0.06% of these comparisons are positioned within intra-cluster boundaries (see below), so this parameter does not substantially impact results. Fewer than 9% of the 2.4 million pairwise comparisons between dsDNA viruses contain nucleotide distances < 0.5 , reflecting the large genetic diversity in the dataset.

2.2.7 Plotting genomic similarity

Genomic similarities were plotted with RStudio (version 0.99.903, implementing R version 3.3.0). Each point represents a single pairwise comparison, and its position in the scatter plot reflects the genomic relationship between the two phages. Sectors of the genomic similarity plot were defined by assessment of the distributions of actinobacteriophages based on their cluster and subcluster designations (see below).

2.2.8 Assigning evolutionary mode

The evolutionary mode for each phage was determined (Supplementary Table 2-1). Each pairwise comparison was classified in the HGCF evolutionary mode if a) the Mash distance was less than 0.16 and gene content dissimilarity was above the line $y = 3.5x$, or b) the Mash distance was greater than 0.16 and gene content dissimilarity was above the line $y = 2x + 0.25$. Otherwise, the pairwise comparison was classified in the LGCF evolutionary mode. To assign evolutionary mode to individual phages, only comparisons that are distributed within intra-cluster boundaries (nucleotide distance < 0.42 , GCD < 0.89) and within the intermediate range of similarity (where they are not positioned within the regions where the two modes converge, reflected by nucleotide distance < 0.06 and GCD < 0.22 or nucleotide distance > 0.28 and GCD > 0.79) were used. For each phage, the proportion of comparisons assigned to the HGCF mode relative to all assignable comparisons were computed. With these thresholds, 60% of phages can unambiguously be assigned to one mode or the other since they exhibit pairwise comparisons that are positioned completely within the one of the two sectors. For the rest of the phages, they were classified as a) “HGCF” if $> 80\%$ of comparisons are distributed in the HGCF mode, b) “LGCF” if $< 20\%$ of comparisons are distributed in the HGCF mode, c) “Mixed” if 20-80% of comparisons are distributed in the HGCF mode, d) “Unknown” if no comparisons were distributed in either of the two mode regions.

Although the thresholds and parameters are intended to predict the evolutionary mode in an unbiased, conservative manner, there are nevertheless phages that exhibit a “Mixed” distribution. This designation is likely an artifact for many of these phages based on the conservative thresholds used. For instance, while 15 of the Subcluster A1 phages are assigned “HGCF”, 40 are assigned “Mixed”, not because they are broadly distributed across the LGCF

sector, but simply due to their distant pairwise relationships with other Cluster A phages that are positioned within close proximity to the conservative boundaries. Classification of individual actinobacteriophage clusters were based on the classification of their constituent phages. Clusters were classified as a) “HGCF” if they contain phages classified as HGCF and contain no phages classified as LGCF or Mixed, b) “LGCF” if they contain phages classified as LGCF and contain no phages classified HGCF or Mixed, c) “Mixed” if more than one phage was classified as Mixed, and d) “Unknown” if all phages were classified as Unknown.

2.2.9 Calculation of genome size disparities

The genome size disparity was computed for each pairwise comparison by determining the absolute difference between the two phage genome sizes, determining the proportion of the size difference relative to each individual genome, and averaging the two proportions. Sliding window averages were then computed by sorting all data points by whole genome nucleotide distance and using the *runmean* function in the *caTools* R package to compute average genome size disparities within sliding windows of 101 data points.

2.2.10 Calculation of shared and unshared gene subset data

Shared and unshared nucleotide distances were computed as follows. Gene nucleotide sequences were extracted from genomes and analyzed using custom python scripts. For each comparison, phams were categorized as “shared” or “unshared” depending on whether they were present in both genomes or only one genome. Gene sequences for both genomes were categorized as “shared” or “unshared” depending on their associated pham’s assigned category.

For each genome, gene sequences in each category were concatenated into a single nucleotide sequence. Each pairwise comparison thus resulted in four concatenated nucleotide sequences. Mash was used with the whole genome optimized parameters to compute nucleotide distances between the shared gene sequences, or between the unshared gene sequences, within each pairwise phage comparison. All shared and unshared nucleotide distances with a p -value $< 1 \times 10^{-10}$ were less than 0.6; therefore, all insignificant data points were set to 0.6 so that all data is retained for each plot.

Each comparison thus resulted in a shared gene nucleotide distance and unshared gene nucleotide distance, in addition to the previously computed whole genome nucleotide distance (Supplementary Table 2-2). Sliding window averages for shared and unshared gene distances were computed similar to the genome size disparities, in which all data points were first sorted by whole genome GCD. The proportion of total coding sequence derived from unshared genes in each genome was computed by dividing the length of the concatenated unshared gene sequences by the combined length of shared and unshared gene sequences (i.e. the genome's total coding sequence). For each comparison, the unshared coding sequence proportions of both genomes were averaged.

For all phams present in clustered actinobacteriophages, a “pham distribution” metric was computed, reflecting the total number of clusters or singletons in which the pham is present at least once. The average pham distribution for shared and unshared phams was computed for each pairwise comparison. The frequency of orphams (phams that contain a single gene) reflects the total number of orphams present for each comparison. Sliding window averages for these two metrics were computed as for genome size disparities.

2.2.11 Analysis of prokaryotic Virus Orthologous Groups

Prokaryotic Virus Orthologous Groups (pVOG) data for 1,877 phages in the *Bacteriophages_2333* dataset were downloaded from the pVOG database (Grazziotin et al., 2017) on 2/21/2017. Pairwise GCD was computed for this subset of phages using VOG data instead of with pham data.

2.2.12 Creating cluster-specific multi-gene phylogenies

To compare Mash-based nucleotide distance to phylogenetic branch lengths, phylogenetic trees were created for specific clusters that differ in lifestyle and evolutionary mode (Table 2-1). The phylogenies are based on several structural and assembly genes, which tend to be the most highly conserved genes (Hatfull and Hendrix, 2011). Highly conserved structural/assembly genes were identified by manually assessing which phams were present in a majority of phages per cluster and assessing their predicted function based on SEA-PHAGES annotations. Multiple genes were used for each phylogeny, and since different clusters tend to be highly unrelated, it is not possible to create phylogenies based on the same exact types of genes.

Table 2-1. Phams used to construct multi-gene phylogenies.

Cluster	Pham	Function	Ave. concatenated length (amino acids)
A	2847	head-to-tail connector	263
	22298	head-to-tail connector	
B	3753	minor tail subunit	2,214
	5322	capsid morphogenesis	
	22085	major capsid	
	22421	major tail subunit	
F	431	head-to-tail connector	1,470
	1120	head-to-tail connector	
	3237	minor tail subunit	
	5414	head-to-tail connector	
	5557	minor tail subunit	
	16649	major tail subunit	
K	2031	major tail subunit	2,203
	4865	head-to-tail connector	
	7777	minor tail subunit	
	8258	major capsid	
	21863	scaffold	
	21994	terminase large subunit	
	22458	portal	
BD	4535	head-to-tail connector	1,654
	5928	scaffold	
	6139	major capsid	
	21799	minor tail subunit	
	21936	major tail subunit	
	22504	portal	

The number of genes used for each cluster-specific phylogeny varied due to the number of highly conserved structural and assembly genes available for analysis in that cluster as well as the average size of the gene (where it was attempted to create concatenated alignments that were comparable in length between clusters). Notably, for Cluster A phages only two genes were used. Originally, phams 7209 (head-to-tail connector) and 22174 (portal) were also used, but a horizontal gene transfer event among three of the Subcluster A1 phages resulted in an unreliable phylogenetic tree for downstream analysis, so these two genes were not used for the final tree. Protein sequences for each gene set were aligned using webPRANK (Loytynoja and Goldman, 2010) using the default settings, all gene alignments for each specific cluster were concatenated,

and the concatenated alignments were used to construct phylogenies using maximum likelihood in SeaView (Gouy et al., 2010) using default settings.

2.2.13 Measuring rates of horizontal gene transfer

Rates of HGT were computed using Count (Csuros, 2010). Using the cluster-specific multigene phylogenies and cluster-specific presence/absence pham tables, Count predicted the gain and loss of individual phams across the phylogeny using Wagner parsimony (with equal penalties for gains and losses). For each branch in the tree, gain/loss events were matched to amino acid distances using the *ete3* python package. The total gain/loss events were divided by the total branch lengths in the tree (or A1 and non-A1 subtrees in the Cluster A analysis) resulting in the HGT rates, similar to what has been done previously in bacterial genomes (Puigbo et al., 2014). Summed gain and loss events for phams in specific gene function categories were divided by total branch lengths to obtain gene function-specific HGT rates. Gene function-specific rate deviation from expected is the log2-transformed ratio of the proportion of the category's HGT rate relative to the total HGT rate in the cluster divided by the proportion of phams in the category relative to all phams in the cluster. It is important to note that when larger gain penalties are used, the absolute number of predicted pham gains decreases while that of pham losses increases, as expected, and the sum total of gains and losses increases. Regardless of gain penalty though, the sum total of gains and losses remains proportionally larger in HGCF phages compared to LGCF phages. Therefore, the results using a gain penalty of 1 are reported, which reflect the smallest sum total of gains and losses predicted and are thus a conservative estimate of horizontal gene transfer rates.

2.2.14 Analysis of LysB horizontal gene transfer

All Cluster F phages contain one of two LysB phams (21902 and 6754). All protein sequences from clustered actinobacteriophages that are assigned to either of these two phams were retrieved, and a single alignment and phylogenetic tree was constructed as for the cluster-specific multigene analysis.

2.2.15 Data analysis

Custom scripts used to process, compute, and analyze evolutionary mode data are available on Github (https://github.com/tmavrich/mavrich_hatfull_nature_micro_2017).

2.2.16 Computing MaxGCDGap

MaxGCDGap analysis was performed on data from 789 actinobacteriophages using the Phamerator database, *Actinobacteriophage_789* [Supplementary Table 2-3, originally published in (Pope et al., 2017)]. MaxGCDGap analysis was also performed on data from 209 phages infecting bacterial hosts in the phylum Cyanobacteria using a separate Phamerator database, *Cyanobacteriophage_209*. Cyanobacteriophage genomes were retrieved from RefSeq or the GenBank *nr* database. To compute MaxGCDGap, GCD for all pairwise comparisons were computed. For each phage, all GCDs involving that phage were ranked by magnitude, and the difference between consecutive GCDs (GCD gap) were computed. A dummy GCD data point of 0 was included in the ranked list to compute the GCD gap between the phage of interest and the nearest relative. GCD gaps were ranked by magnitude such that the largest value represents the

maximum gap (MaxGCDGap), which ranges from near 0 (indicating small GCD discontinuities) to 1 (indicating large GCD discontinuities). Box plot distributions of MaxGCDGap were plotted in R. Custom python and RStudio scripts used to process data are available on Github (https://github.com/tmavrich/pope_mbio_2017).

2.3 RESULTS

2.3.1 Generation of genomic similarity plots

To examine phage evolutionary patterns, I created a Phamerator database consisting of 2,191 dsDNA phages isolated and sequenced from the SEA-PHAGES program as well as phages stored in the NCBI RefSeq database (Supplementary Table 2-1). The phages are diverse and infect bacterial hosts from 9 phyla and 75 families. Several types of viruses other than dsDNA phages were also included as negative controls for the evolutionary analyses (see Materials and Methods).

Examination of broad patterns of phage evolution and horizontal gene transfer (HGT) requires quantitative metrics to directly compare all phages. However, since phages are genetically diverse, there are no completely conserved phage genes that facilitate universal phage comparisons analogous to the 16S rRNA gene used for bacterial comparative genomics (Lima-Mendez et al., 2008; Rohwer and Edwards, 2002). Instead, phages have been qualitatively categorized based on assessment of their morphological features (such as head and tail structure), genomic features (such as gene content and nucleotide sequence), or environmental features (such as bacterial host or source of isolation)(Hatfull et al., 2010; Lawrence et al., 2002; Lima-

Mendez et al., 2008; Maniloff and Ackermann, 1998; Proux et al., 2002; Rohwer and Edwards, 2002). Therefore, I directly compared all phages using a bivariate “genomic similarity” plot reflecting whole genome gene content dissimilarity (GCD) and whole genome nucleotide distance, similar to previous bacterial and archaeal studies (Tu and Lin, 2016; Varghese et al., 2015; Wolf et al., 2016).

Computing whole genome GCD and nucleotide distance for the 2.7 million viral pairwise comparisons is not straightforward though. Identifying nucleotide or amino acid sequence similarities using the most rigorous, alignment-based tools such as BLAST (Altschul et al., 1990) and average nucleotide identity (ANI)(Konstantinidis and Tiedje, 2005) are computationally expensive such that it is prohibitive to use them for this dataset. Recently, several tools have become available that facilitate alignment-free sequence comparisons within large datasets by analyzing and comparing subsets of short sequences, or “kmers”, including kClust (Hauser et al., 2013), FFP (Sims et al., 2009), and Mash (Ondov et al., 2016). Therefore, I utilized these alignment-free, kmer-based tools that can approximate alignment-based sequence similarity so that I can quantify these two metrics for all pairwise comparisons.

Differences in actinobacteriophage gene content is typically measured using the Phamerator pipeline (Appendix A), which utilizes kClust (Hauser et al., 2013) to group similar gene products into phamilies (“phams”) of related protein sequences. The alignment-free approach utilizes kmers to measure amino acid sequence similarity, and parameters can be adjusted to group genes similar to alignment-based strategies such as BLAST (Altschul et al., 1990) to reflect homology and shared evolutionary history. Charles Bowman implemented and optimized the kClust parameters for the actinobacteriophage phages in the `k_phamerate.py` script

(Appendix A). The ~ 230,000 genes from the 2,333 viral genomes have been grouped into 62,363 phams, of which 55% are orphams (i.e. the phams contain a single gene).

The degree of sequence diversity and HGT reflected by pham designation can be illustrated with LysB proteins. Many phages carry *lysB* genes, and they exhibit substantial sequence diversity (Payne and Hatfull, 2012). In the current database, over 80% of Cluster F phages encode a LysB protein in pham 21902 while the rest encode a LysB protein in pham 6754. For example, Cluster F phages Fruitloop and Ovechkin exhibit substantial sequence similarity across their genomes, and contain *lysB* genes, gene 30 and gene 32 respectively, at syntenic positions (Figure 2-1A). However, there is no significant nucleotide sequence similarity across this locus, and Fruitloop gp30 has been grouped into pham 21902 while Ovechkin gp32 has been grouped into pham 6754. Corndog is not genetically related to Fruitloop and is grouped into Cluster O, but Corndog gene 70 and Fruitloop gene 30 exhibit some degree of nucleotide sequence similarity and their gene products have been grouped into the same LysB pham, 21902 (Figure 2-1A). There are 149 genes present in phages spanning six clusters and one singleton genome that have been grouped into either pham 21902 or 6754. A phylogenetic tree derived from an alignment of all genes in these two phams is consistent with the pham designations (Figure 2-1B). All genes in pham 21902 represent a monophyletic clade distinct from taxa in pham 6754, regardless of the phage cluster. Furthermore, phages in Clusters A and L encode homologs of LysB in phams 21902 and 6754, but they also encode LysB proteins from other phams as well, indicating the patterns observed for LysB phams in Fruitloop, Ovechkin and Corndog are not unique (Figure 2-1C).

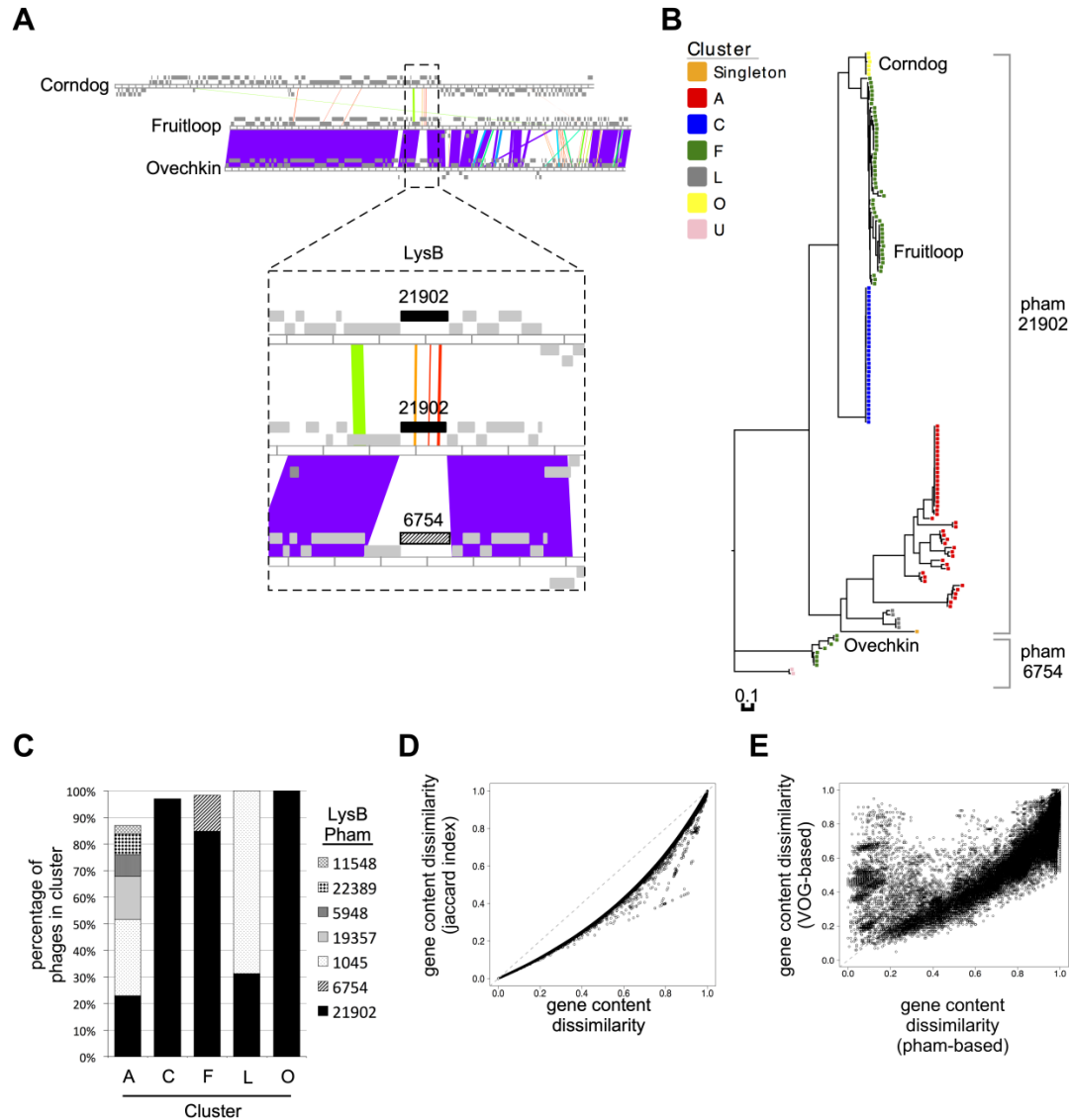


Figure 2-1. Comparison of gene content derived from kmer-based and alignment-based tools.

(A) (top) BLAST-based whole genome alignments in Phamerator of two Cluster F phages (Fruitloop, Ovechkin) and the Cluster O phage, Corndog, harboring *lysB* genes. Spectrum color shading reflects BLASTN *e*-value significance of aligned regions, ranging from unrelated (white) to closely related (violet). (bottom) Enlarged view of the *lysB* locus, with phams indicated. (B) Phylogeny constructed from alignment of all LysB amino acid sequences grouped into phams 21902 and 6754 and present in actinobacteriophages. Taxa are represented by boxes colored according to assigned cluster with Corndog, Fruitloop, and Ovechkin specifically highlighted. (C) For phage clusters that encode at least one LysB protein in pham 21902, all LysB phams were identified. Bar graphs report the percentage of phages in each cluster that encode a LysB protein in each indicated pham, (D) Scatter plot comparing pham-based gene content dissimilarity and Jaccard indices for all viral comparisons. (E) Scatter plot comparing pham-based and VOG-based gene content dissimilarities. Figure adapted from (Mavrich and Hatfull, 2017).

To examine pairwise gene content diversity between all phages, I used a dissimilarity index, where larger values indicate greater dissimilarity, similar to the Jaccard (Lima-Mendez et al., 2008) and Bray-Curtis (Tu and Lin, 2016) gene content dissimilarity indices used in other evolutionary studies. For each pairwise comparison, gene content dissimilarity (GCD) reflects the degree to which each phage's set of phams differ from each other (see Materials and Methods). This GCD index is computed similarly to the more common Jaccard index, and the two dissimilarity values for all ~ 2.7 million pairwise viral comparisons are highly correlated (Figure 2-1D).

Using GCD, I compared the pham creation strategy in Phamerator to the alignment-based gene grouping strategy used to create prokaryotic viral clusters of orthologous genes (VOGs)(Grazziotin et al., 2017). GCD derived from phams and VOGs are highly correlated (Figure 2-1E). Some discrepancy is observed, but this is expected. In Phamerator, all genes are assigned a pham, even if they are the sole members of the pham (i.e. orphans). In contrast, the strategy to create VOGs relies not only on sequence alignment parameters but ortholog abundance: genes that lack orthologs in more than two other genomes are not assigned a VOG (Tatusov et al., 1997). As a result, on average only ~ 60-70% of proteins per phage genome are assigned to VOGs (Grazziotin et al., 2017), and this would increase VOG-based GCD scores relative to pham-based GCD scores.

Changes in whole genome gene content can be directly compared to changes in whole genome nucleotide sequence. Mash provides a rapid, alignment-free strategy to measure nucleotide sequence similarity between genomes using a kmer-based strategy. The software computes a nucleotide distance score in which larger values indicate greater degrees of dissimilarity using several user-adjusted parameters. The primary parameters, kmer size and

sketch size, define the length and number of kmers used in the comparison, respectively, and can be adjusted such that the kmer-based distance scores can highly correlate with alignment-based ANI scores. Guidelines are provided to cluster sequences at the species level (Ondov et al., 2016), where ANI is typically < 0.05 , but these are not sufficient to investigate broader evolutionary patterns. The correlation of Mash distances to ANI distances varies based on sequence size and similarity; larger sketch sizes and smaller kmer sizes provide more accurate estimates between genomes of greater distance, but larger kmer sizes provide more accurate estimates between larger genomes. Therefore, I optimized Mash distances such that they approximate ANI over a broad sequence span using a training set of 79 phage genomes representing multiple host phyla and actinobacteriophage clusters.

Mash distances were computed from a matrix of kmer and sketch size combinations. All distances with a p -value $< 1 \times 10^{-10}$ were retained and compared to distances derived from ANI (Figure 2-2A). In general, Mash distances correlate well with ANI when sketch sizes are greater than 1,000; additionally, as sketch sizes increase and kmer sizes decrease, a greater amount of statistically significant Mash distances is plotted. From this analysis, a kmer size of 15 and a sketch size of 25,000 were chosen, in which Mash and ANI distances correlate up to a Mash distance of ~ 0.33 (ANI distance of ~ 0.4). Additionally, since phage genomes in this dataset range in size from 2.5 kb to nearly 500 kb, I added a parameter to minimize potential distortions generated from large genome size disparities. I limited the analysis to comparisons between genomes that had a maximum genome size disparity of 100%, ensuring that for each pairwise comparison one genome was no more than twice the size of the other genome (Figure 2-2B).

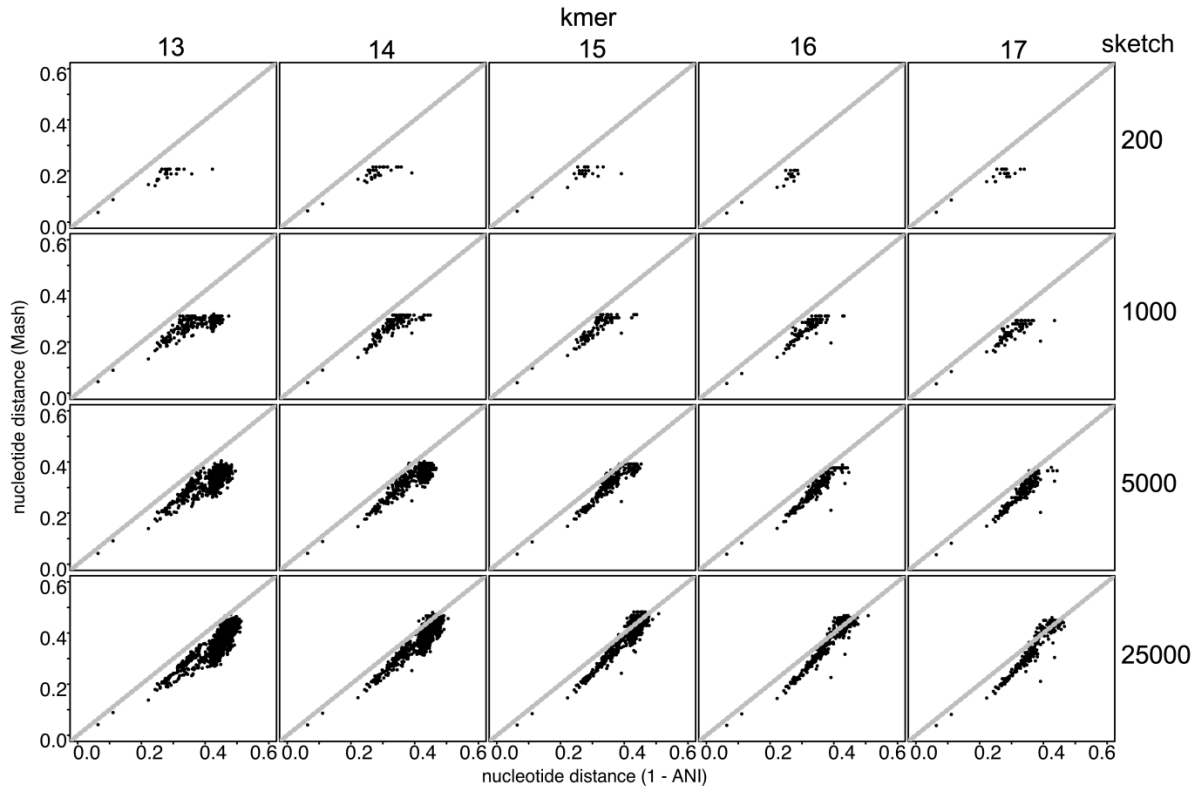
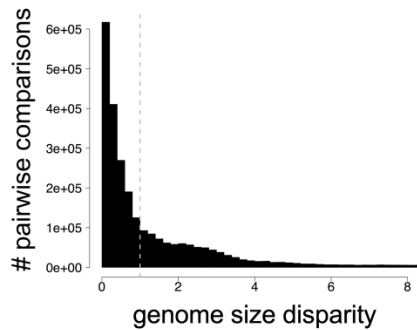
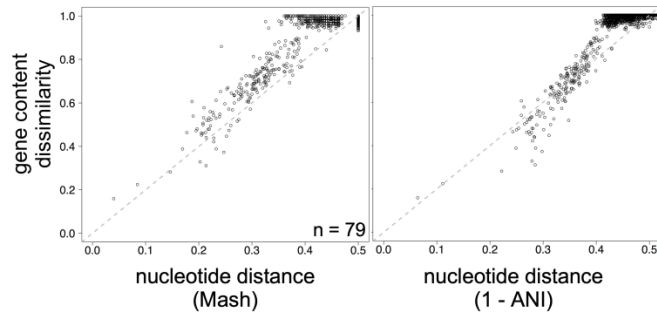
A**B****C**

Figure 2-2. Optimization of Mash to compute kmer-based nucleotide distance.

(A) Mash nucleotide distances among 79 phages of diverse subcluster, cluster, and host phyla were computed with a range of kmers (13 to 17) and sketch sizes (200, 1000, 5000, 25000) and compared to ANI-based distance (1 – ANI). Each point represents an individual pairwise comparison. The $y = x$ line is plotted for reference. Only data points with p -values $< 1 \times 10^{-10}$ and distances < 0.6 are plotted to highlight the correlated metrics. (B) Histogram reflecting the distribution of genome size disparities for all 2.7 million pairwise comparisons in the dataset. The dashed line indicates a genome size disparity of 100%, used as a threshold for comparisons. (C) Genomic similarity scatter plots, indicating genomic relationships for the 79 training set phages, using (left) Mash-based or (right) ANI-based nucleotide distances. Figure adapted from (Mavrich and Hatfull, 2017).

The bivariate genomic similarity plot comparing whole genome nucleotide distance and gene content dissimilarity should be able to reflect the genomic relationship between any pair of phages. Mash distances for all 2.7 million comparisons in the database range from 0 (identical sequence) to 1 (no similarity). However, no statistically significant Mash score was greater than 0.5, and statistically insignificant data had been removed for Mash optimization (Figure 2-1A). Therefore, in order to retain these data in the genomic similarity plot, all distances that are greater than 0.5 or that have a p -value $> 1 \times 10^{-10}$ are converted to 0.5. Using this strategy, genomic similarity plots depict the relationship between nucleotide distance (ranging from 0, complete identity, to 0.5, unrelated) and gene content dissimilarity (ranging from 0, identical phams, to 1, no identical phams) for any pair of phages (Figure 2-2C). Data points positioned towards the top right of the plot indicate phage pairs with no nucleotide or gene content similarity, and data points positioned towards the bottom left of the plot indicate phage pairs with high nucleotide and gene content similarity. Genomic similarity plots for all phage pairs in the training set indicate that similar correlations between changes in gene content and nucleotide distance are observed when Mash distances or ANI distances are used (Figure 2-2C).

2.3.2 Phages exhibit two evolutionary modes

Mosaicism has been observed in phages across nearly every host phyla, and genomic similarity plots provide insight into general and specific patterns of phage evolution. Using the optimized parameters, genomic similarities between all 2.4 million pairwise comparisons involving the 2,191 dsDNA bacteriophages in the dataset were computed (Figure 2-3A, Supplementary Table 2-2). The majority of pairwise comparisons are positioned near the top right of the plot, reflecting the diversity among the phages. The remaining genomic relationships

are distributed across the plot, reflecting that phages in this dataset exhibit a genetic spectrum, similar to previous reports (Pope et al., 2015). Furthermore, these relationships form two distinct distributions, suggesting there may be different rates of change in gene content (gene content flux) compared to nucleotide sequence (Figure 2-3).

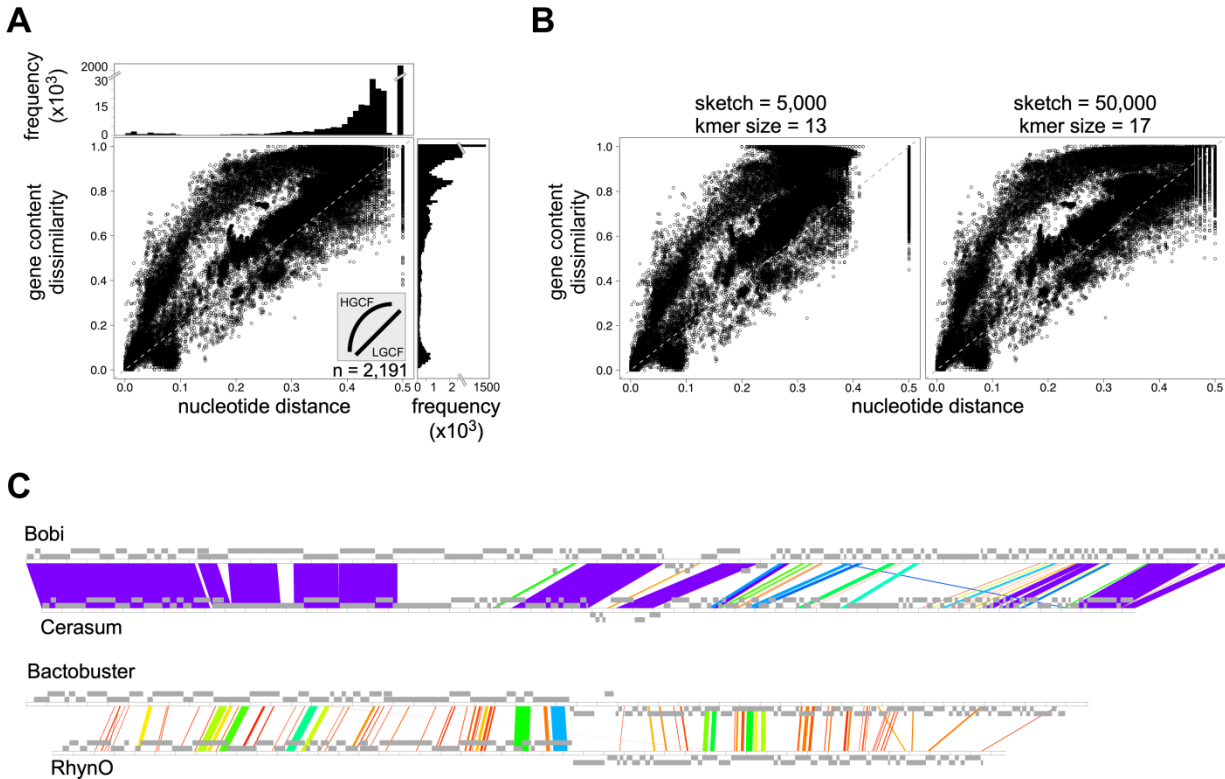


Figure 2-3. Phages exhibit two evolutionary modes.

(A) Genomic similarity scatter plot comparing Mash-based nucleotide distance and pham-based gene content dissimilarity for 2.4×10^6 dsDNA phage comparisons. The line at $y = 2x$ is plotted for reference. Marginal frequency histograms emphasize densely-plotted regions, with truncated y axes for viewability. (inset) Diagram defining distributions for HGCF and LGCF evolutionary modes. n = number of dsDNA phages used for the analysis. (B) Genomic similarity scatter plots as in panel A for dsDNA phages using different Mash sketch size and kmer parameters to compute nucleotide distance. (C) BLAST-based whole genome alignments in Phamerator of Bobi and Cerasum (Cluster F phages) and Bactobuster and Rhyno (Cluster A phages) representing HGCF and LGCF evolutionary modes, respectively. Both comparisons have approximately equal gene content dissimilarities (0.51 and 0.50, respectively), but markedly unequal whole genome nucleotide distances (0.07 and 0.25, respectively). Spectrum color shading as in Figure 2-1A. Figure adapted from (Mavrich and Hatfull, 2017).

I investigated whether the two distributions are artifacts of the Mash distance metric. The same genomic relationships are observed when either less stringent parameters (sketch size = 5,000, kmer size = 13) or more stringent parameters (sketch size = 50,000, kmer size = 17) are used to compute nucleotide distance (Figure 2-3B). Additionally, the two modes correlate with BLAST nucleotide sequence alignments, such as between Cluster F phages Bobi and Cerasum and between Cluster A phages Bactobuster and Rhyno (Figure 2-3C). Both comparisons exhibit GCD ~ 0.5 , so they have comparable differences in gene content. Bobi and Cerasum genomes, which have a Mash nucleotide distance of 0.07 and are positioned in the HGCF mode, exhibit substantial BLAST nucleotide sequence similarity across their genomes. In contrast, Bactobuster and Rhyno genomes, which have a Mash nucleotide distance of 0.25 and are positioned in the LGCF mode, exhibit much lower BLAST nucleotide sequence similarity across their genomes. Therefore, the two distributions are likely to have a biological basis and suggest phages may evolve in two distinct ways (“modes”), designated here as high (HGCF) and low (LGCF) gene content flux (Figure 2-3A).

To inform my interpretation of the genomic similarity plot, I compared relationships between different groups of viruses, including the 142 viruses that are not dsDNA bacteriophages. Viruses of different host domains and nucleic acid genomes are expected to be unrelated (Lawrence et al., 2002; Lima-Mendez et al., 2008; Roux et al., 2015). In my analysis, all comparisons involving a bacteriophage and the eukaryotic virus, *Tetrasetmis viridis* virus S1, exhibit nucleotide distances of 0.5 and GCD of 1 (i.e. no measurable similarity). The majority of genomic relationships between bacteriophages and the three archaeal viruses exhibit nucleotide distances greater than 0.38 and GCD of 1 (Figure 2-4A). Similarly, the majority of genomic relationships between bacteriophages of different nucleic acid types (dsDNA, ssDNA, dsRNA,

ssRNA) exhibit nucleotide distances greater than 0.4 and GCD greater than 0.8 (Figure 2-4B). Over 99% of comparisons between dsDNA phages of different host phyla exhibit either nucleotide distances greater than 0.3 and GCD greater than 0.95 (Figure 2-4C). Viruses with ssDNA, dsRNA, or ssRNA genomes are not abundant enough to determine if they exhibit similar evolutionary modes as viruses with dsDNA genomes (Figure 2-4D).

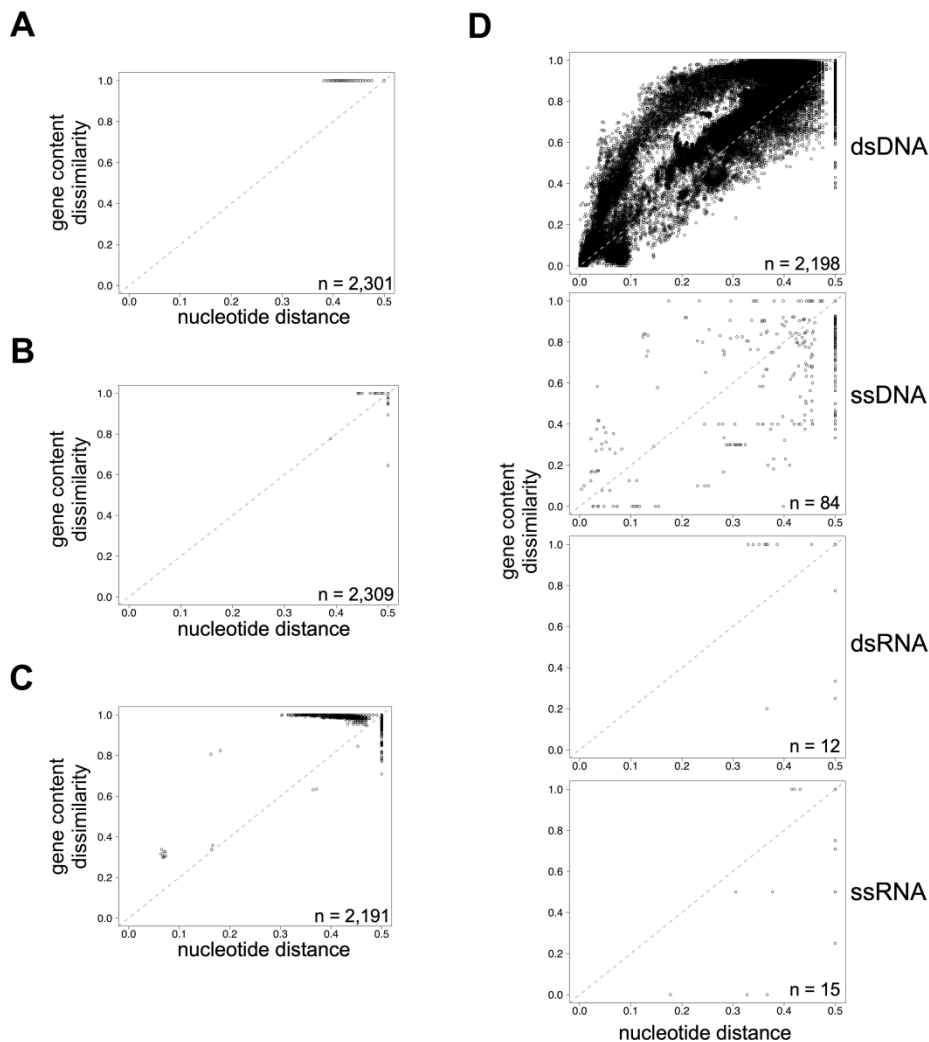


Figure 2-4. Evaluation of genomic relationships between different types of viruses.

Genomic similarity scatter plots involving (A) one archaeal virus and one bacteriophage, (B) viruses of different nucleic acid types (dsDNA, ssDNA, dsRNA, ssRNA), (C) dsDNA phages of different host phyla, and (D) viruses with the same type of nucleic acid. n = number of viruses used for each analysis. Data in all panels plotted as in Figure 2-3A. Figure adapted from (Mavrich and Hatfull, 2017).

Next, I compared and contrasted genomic relationships specifically among actinobacteriophages. These phages have been manually annotated through the SEA-PHAGES program and have been systematically grouped into clusters that reflect degrees of genetic relationships based on qualitative and quantitative assessment of gene content and nucleotide similarity (Supplementary Table 2-1)(Hatfull et al., 2010). Nearly 100% of “intra-cluster” comparisons involving phages within the same cluster exhibit nucleotide distances below 0.42 and gene content dissimilarities below 0.89 (Figure 2-5A). Clustered phages are further subdivided into subclusters if more substantial genetic relationships are apparent, and nearly 100% of “intra-subcluster” comparisons involving phages within the same subcluster exhibit nucleotide distances below 0.2 and gene content dissimilarities below 0.62 (Figure 2-5A). Many clustered phages are not further subdivided due to insufficient genetic diversity within the cluster, and the majority of intra-cluster genomic relationships among these types of phages are distributed similarly to the genomic relationships among intra-subcluster comparisons (Figure 2-5A, B). Phages that do not exhibit substantial genomic relationships with any other phages are not clustered and remain as singletons. The majority of genomic relationships between singletons and other phages in the database are distributed similarly to the inter-cluster genomic relationships, exhibiting nucleotide distance greater than 0.2 and GCD greater than 0.5 (Figure 2-5A, C). From these analyses, there is a strong correlation between the SEA-PHAGES clustering strategy and the genomic relationships apparent in the genomic similarity plot.

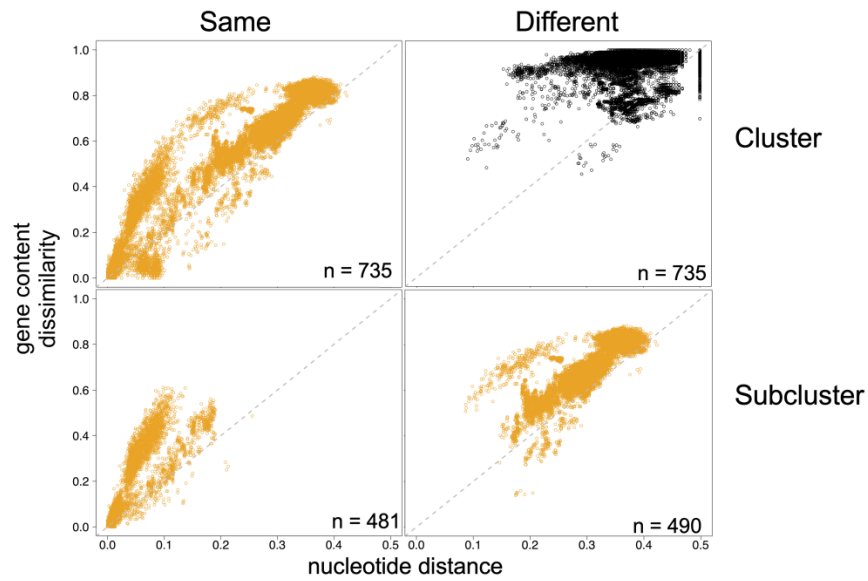
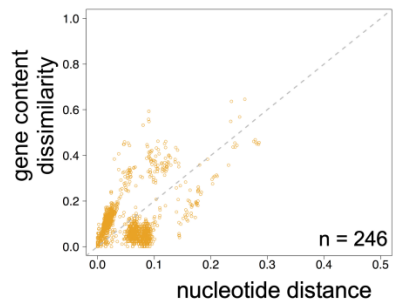
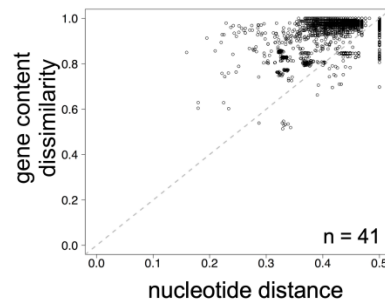
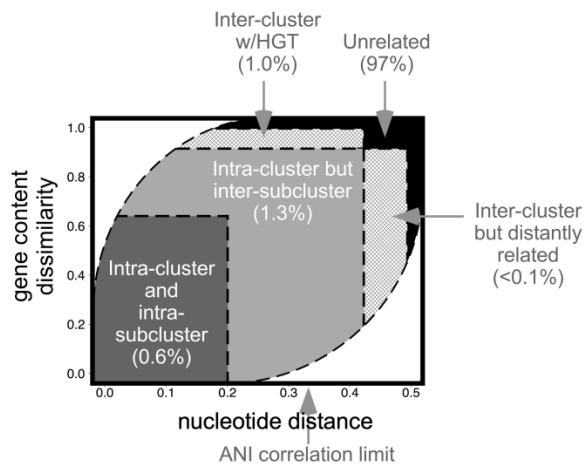
A**B****C****D**

Figure 2-5. Evaluation of genomic relationships between actinobacteriophages.

(A-C) Cluster-specific intra-cluster (orange) and inter-cluster (black) comparisons are plotted as in Figure 2-3A for actinobacteriophages that have been manually clustered by the SEA-PHAGES program. (A) Comparisons are separated by whether they involve phages of the (left) same or (right) different (top) cluster or (bottom) subcluster. (B) Comparisons involving phages of the same cluster that contain no subcluster divisions are plotted. (C) Comparisons involving at least one manually-curated singleton phage are plotted. For panels A and B, n = number of phages used for the analysis. For panel C, n = number of singletons used for the analysis. (D) Defined sectors in the genomic similarity plot (dashed lines), highlighting various genomic relationships (see Materials and Methods), including the percentage of dsDNA phage comparisons in Figure 2-3A positioned in each sector. Figure adapted from (Mavrich and Hatfull, 2017).

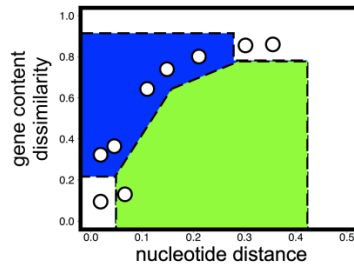
Using the correlations observed with SEA-PHAGES data, different types of genomic relationships in the genomic similarity plot can be defined. Genomic relationships in the “Intra-cluster and intra-subcluster” sector involve phages with enough similarity to be grouped into the same subcluster. Genomic relationships in the “Intra-cluster but inter-subcluster” sector involve phages with enough similarity to be grouped into the same cluster. Genomic relationships in the “Inter-cluster but distantly related” sector involve phages that may lack enough similarity to be grouped into the same cluster, but which may be evolutionary related due to their low GCD. Genomic relationships in the “Inter-cluster w/HGT” sector involve phages that lack enough similarity to be grouped into the same cluster but exhibit low nucleotide distance, suggesting that they have experienced substantial, recent, horizontal gene transfer that skews the whole genome nucleotide distance. Lastly, genomic relationships in the “Unrelated” sector involve phages that lack enough similarity to be grouped into the same cluster, that do not exhibit statistically significant sequence similarity, or that contain large genome size disparities. Using these genomic similarity boundaries, almost 100% of comparisons involving dsDNA phages of different host phyla are positioned in the unrelated sector (Figure 2-4C). Less than 1% of inter-cluster comparisons and about 1% of comparisons involving at least one singleton are positioned

within the intra-cluster sector (Figure 2-5A, C). Fewer than 3.5% of inter-subcluster comparisons between phages of the same cluster and nearly 99% of comparisons involving phages of the same cluster but have not yet been subclustered are positioned within the intra-subcluster sector (Figure 2-5A, B). Among the 2.7 million comparisons involving viral genomes in the entire dataset, 97% are positioned in the unrelated sector highlighting their genetic diversity (Figure 2-5D).

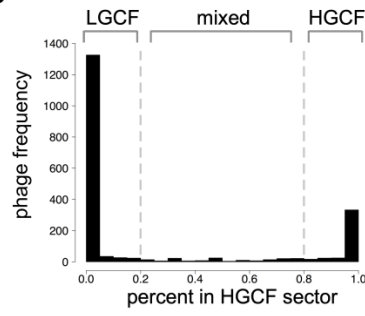
2.3.3 Genetically related phages exhibit specific evolutionary modes

The two evolutionary modes may reflect different degrees of gene content flux that are not associated with the evolution of individual phages. In contrast, individual phages may be constrained to one mode or the other. To examine this correlation, HGCF and LGCF sectors were defined on the genomic similarity plot, and individual phages were assigned an evolutionary mode (HGCF, LGCF, Mixed, or Unknown) based on the frequency of genomic relationships distributed within each sector (Figure 2-6A)(see Materials and Methods). The majority of phages can be easily assigned to either the HGCF or LGCF mode, and very few are designated as Mixed, indicating that individual phages predominantly evolve within only one of the evolutionary modes (Figure 2-6B, Supplementary Table 2-1).

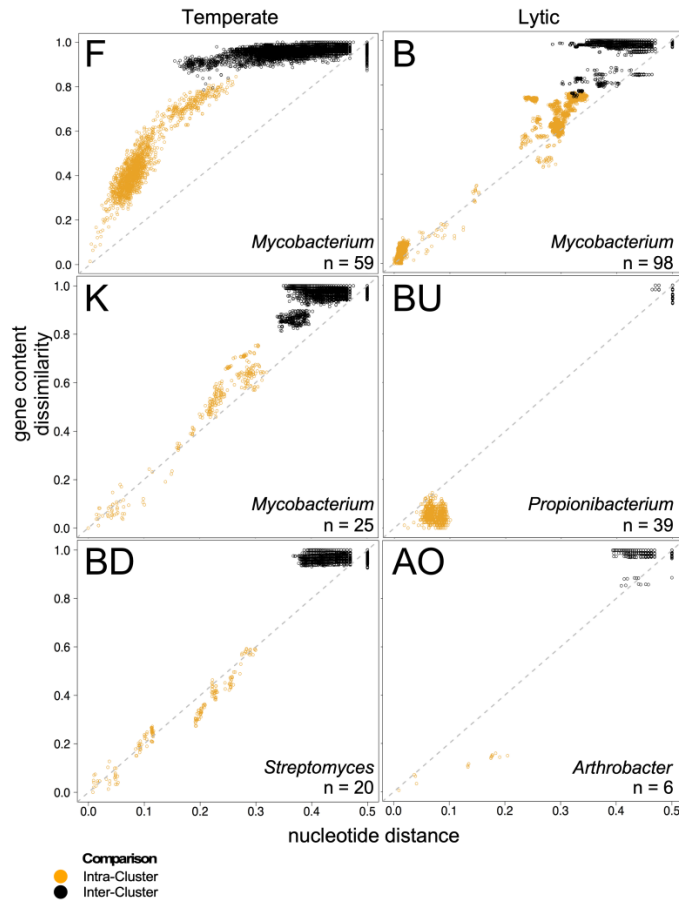
A



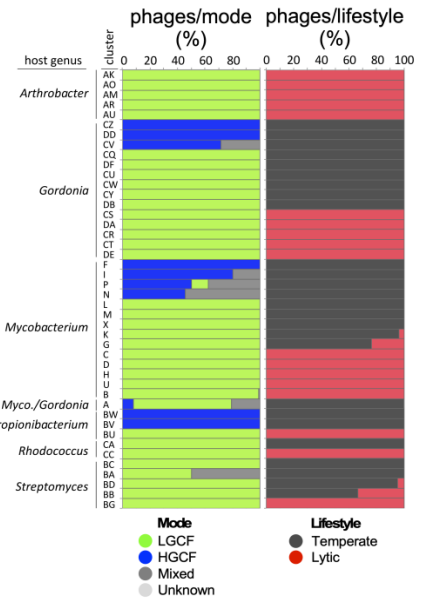
B



C



D



E

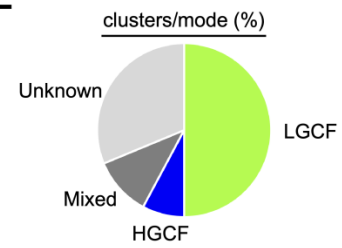


Figure 2-6. Phages and phage clusters exhibit unique evolutionary trajectories.

(A) Individual phages are classified into evolutionary modes with a simple strategy using defined HGCF (blue), LGCF (green), and non-classifiable (white) sectors on the plot. Mock data points (circles) illustrate how the proportion of comparisons in each sector can be used to assign modes (see Materials and Methods). (B) For each phage, the percentage of comparisons in the HGCF mode was computed, and the histogram reflects the frequency distribution of phages across the range of proportions. Dashed lines and brackets indicate the quantitative boundaries to classify phages into the LGCF, mixed, and HGCF modes. (C) Cluster-specific intra-cluster (orange) and inter-cluster (black) comparisons are plotted as in Figure 2-5 for representative actinobacteriophage clusters and grouped by their predicted or known lifestyle, with cluster and host genus indicated. n = number of phages present in the specific cluster. (D) Stacked horizontal bar graphs for 44 actinobacteriophage clusters in which the evolutionary mode of their constituent phages could be determined, along with their predicted lifestyle. For each cluster, the percentage of the constituent phages that are predicted to be temperate or lytic along with the percentage of the constituent phages that are predicted to be in each evolutionary mode, are indicated. (E) Pie chart reflecting the proportion of all actinobacteriophage clusters in each mode (same color scheme as in panel D). Figure adapted from (Mavrich and Hatfull, 2017).

Similarly, genetically related phages predominantly evolve within the same evolutionary mode. For example, phages in Cluster F are distributed within the HGCF mode, and phages in Clusters B, K, and BD are distributed within the LGCF mode (Figure 2-6C). Although phages in Clusters BU and AO exhibit less genomic diversity than other clusters, they can nevertheless be assigned to the LGCF mode (Figure 2-6C). The majority of clusters contain phages that are predominantly in one mode (Figure 2-6D). Clusters were therefore assigned an evolutionary mode, if possible (see Materials and Methods). Of all actinobacteriophage clusters, nearly 50% could be assigned to the LGCF mode, only a few could be assigned to the HGCF mode, and a third could not be assigned to either mode due to insufficient genetic diversity (Unknown) or due to phages failing to unambiguously exhibit one of the two modes (Mixed) (Figure 2-6E). Thus, genetically related phages tend to exhibit one evolutionary mode, and most exhibit the LGCF mode.

2.3.4 Evolutionary modes are correlated with phage lifestyles

The two evolutionary modes may be associated with phage lifestyle, since lifestyle has been associated with other phage evolutionary patterns. Temperate and lytic phages may be subject to different evolutionary constraints (Chen and Lu, 2002; Chithambaram et al., 2014a; Chopin et al., 2001; Popa et al., 2017). Gene content analysis suggested that temperate phages facilitate transfer of genes among the phage population (Dobbins et al., 2004). Codon usage is impacted by phage lifestyle (Chithambaram et al., 2014a; Lucks et al., 2008). *Escherichia* and *Salmonella* temperate phages are enriched for DNA motifs specific to their hosts that may enhance prophage stability and there appears to be selection for the site of integration and orientation relative to host genomic architecture (Bobay et al., 2013). Temperate phages can improve host fitness (Brussow et al., 2004), including providing genetic resources for defense against other bacteria and phages (Bondy-Denomy et al., 2016; Montgomery et al., 2019).

To investigate whether phage lifestyle impacts genomic similarity patterns, genomic similarity plots were created for over 1,000 empirically determined temperate or lytic phages (Figure 2-7A, Supplementary Table 2-1)(see Materials and Methods). Genomic relationships between lytic phages are predominantly distributed within the LGCF mode, such as phages in Clusters B, BU, and AO (Figures 2-6C, 2-7A). In contrast, genomic relationships between temperate phages are distributed within the LGCF mode, such as phages in Clusters K and BD, or within the HGCF mode, such as phages in Cluster F (Figures 2-6C, 2-7A). The distinctly different distributions of genomic relationships are also observed using VOG-based GCD, indicating that they are not an artifact of kmer-based gene clustering (Figure 2-7B). Since empirically determined lifestyle data is available for only ~ 1,000 viruses, lifestyle was bioinformatically predicted for all genomes in the database (see Materials and Methods).

Predicted and empirical lifestyles are in strong agreement (see Materials and Methods). The majority of phages are categorized as lytic (Figure 2-7C). The majority of actinobacteriophage clusters have phages with the same predicted lifestyle (Figure 2-6D). The same genomic relationships are observed based on predicted lifestyle (Figure 2-7D). The majority of lytic phages are assigned to the LGCF mode, with very few assigned to the HGCF mode (Figure 2-7E). In contrast, only half of the temperate phages that can be unambiguously assigned to a mode are designated as LGCF (Figure 2-7E). The evolutionary patterns associated with lifestyle and cluster are also observed using ANI-based nucleotide distance, indicating they are not artifacts of Mash (Figure 2-7F). Thus, the distinctly different distributions suggest that unlike lytic phages, there are two classes of temperate phages. Temperate phages in Class 1 exhibit HGCF and temperate phages in Class 2 exhibit LGCF (Figure 2-7E).

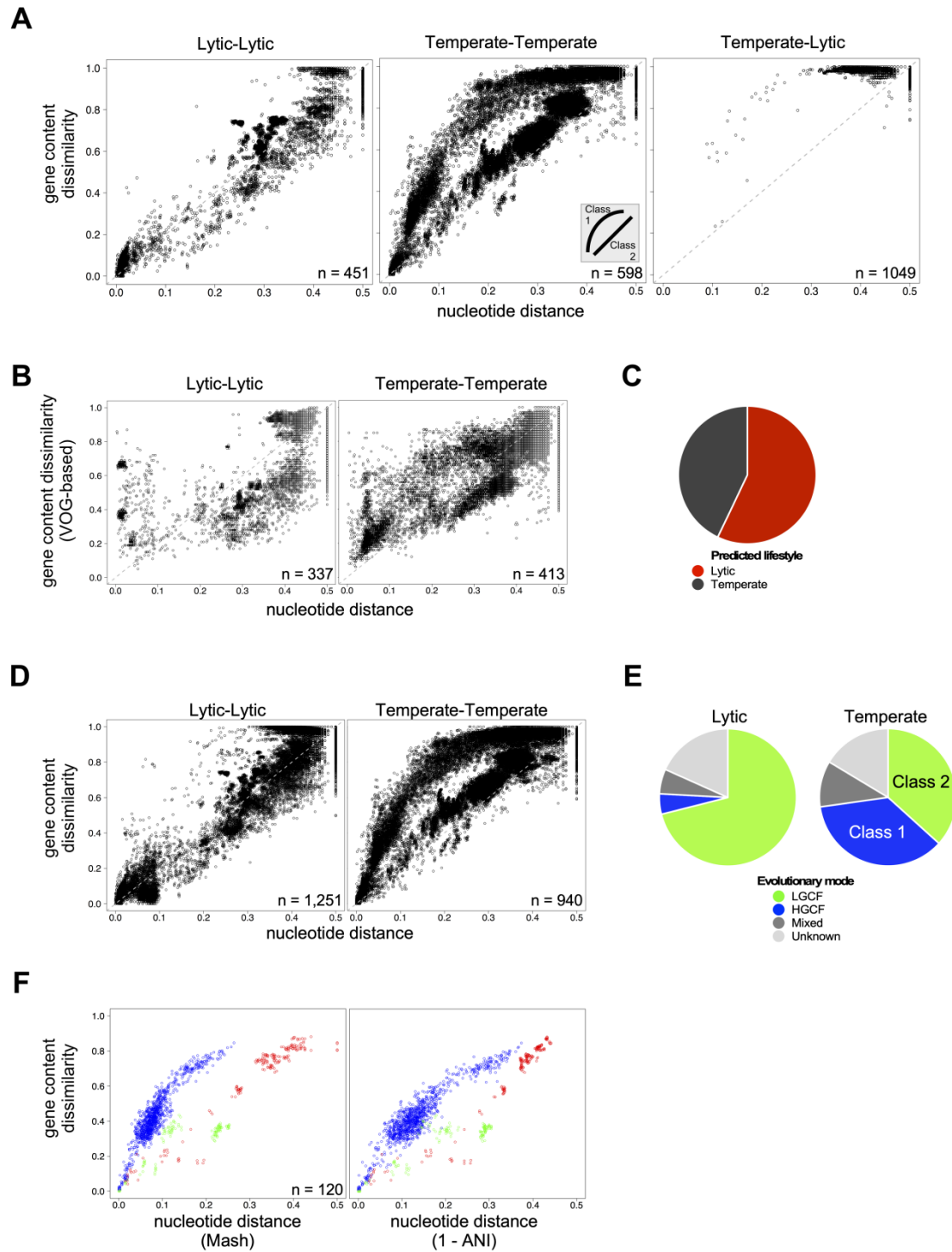


Figure 2-7. Evolutionary modes differ by phage lifestyle.

(A) Genomic similarity scatter plots as in Figure 2-3A involving two lytic (left), two temperate (middle), or one lytic and one temperate phage (right). The middle plot inset defines temperate phage classes for each evolutionary mode. **(B)** Genomic similarity scatter plots as in Figure 2-3A based on empirical phage lifestyle data and VOG-based gene content dissimilarities. n = number of phages used for the analysis. **(C)** Pie chart reflecting the proportion of all phages in the dataset predicted to be temperate or lytic. **(D)** Genomic similarity scatter plots as in Figure 2-3A using all predicted lifestyle data. n = number of phages of each predicted lifestyle used for the analysis. **(E)** Pie charts reflecting the proportion of predicted (left) lytic and (right) temperate phages in the dataset that exhibit the HGCF or LGCF mode. Temperate phages are classified as Class 1 and Class 2 based on evolutionary mode. **(F)** Intra-cluster genomic similarity scatter plots of phages from Clusters F, G, J, L, and N, which represent temperate HGCF (blue), temperate LGCF (green), and lytic (red) groups, using (left) Mash-based nucleotide distances and (right) ANI-based nucleotide distances. Figure adapted from (Mavrich and Hatfull, 2017).

Evolutionary mode may further correlate with other phage characteristics, such as tail structure. Phages can be categorized based on their tail structure as siphoviral (long, non-contractile tail), myoviral (long, contractile tail), or podoviral (short, non-contractile tail), which impacts the types of genes associated with phage structure and assembly (Supplementary Table 2-1)(Krupovic et al., 2011). Siphoviral and myoviral temperate phages may exhibit either evolutionary mode, whereas podoviral temperate phages are predominantly in the HGCF mode (Figure 2-8A). Therefore, evolutionary modes are not strictly associated with a particular tail type.

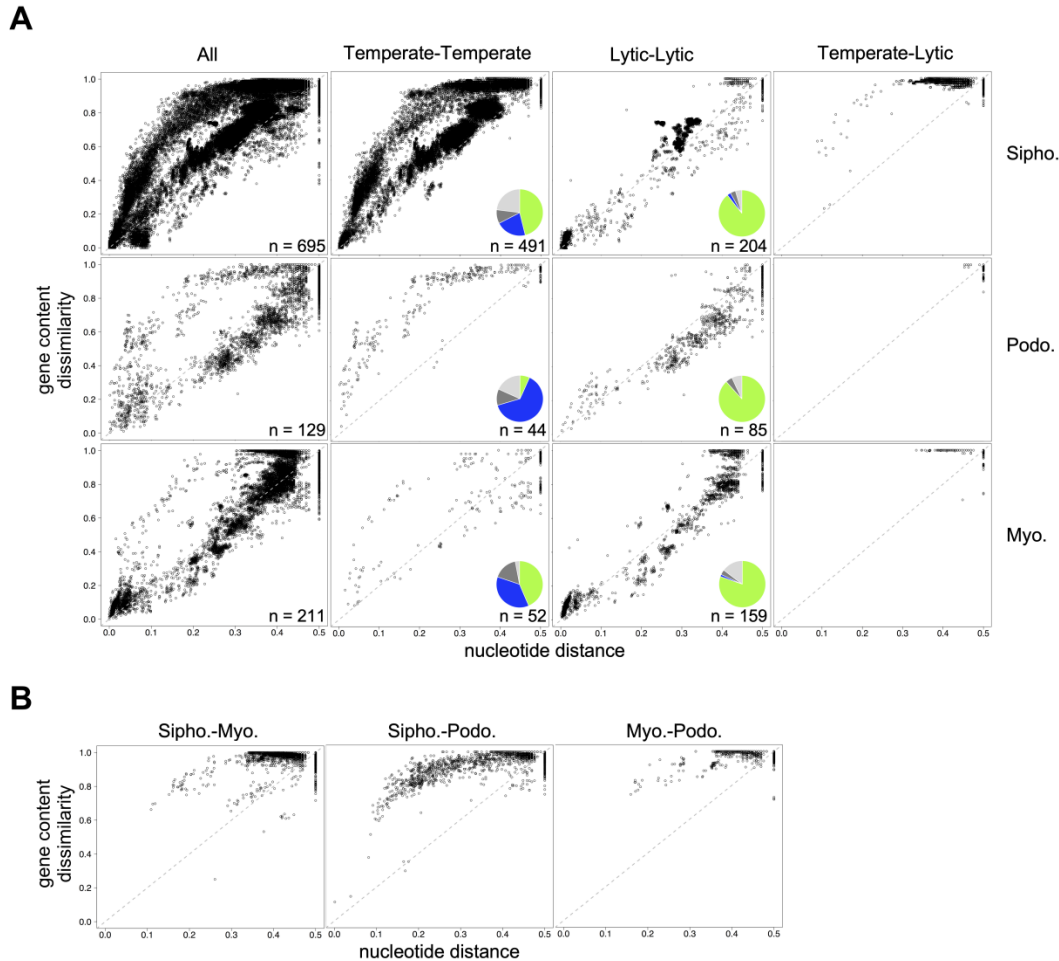


Figure 2-8. Evolutionary modes are not correlated with phage tail morphotype.

(A) Genomic similarity scatter plots as in Figure 2-3A based on phage tail morphotype: *Myoviridae* (Myo), *Siphoviridae* (Sipho), and *Podoviridae* (Podo). Comparisons are plotted involving two phages regardless of lifestyle (All), two temperate phages (Temperate-Temperate), two lytic phages (Lytic-Lytic), or one temperate and one lytic phage (Temperate-Lytic). Pie charts reflect proportion of phages in each evolutionary mode, as in Figure 2-7E. n = number of phages used for each analysis. **(B)** Genomic similarity scatter plots as in Figure 2-3A involving dsDNA phages of different tail morphotypes regardless of phage lifestyle. Figure adapted from (Mavrich and Hatfull, 2017).

2.3.5 Evolutionary modes are correlated with HGT

Two evolutionary modes may be caused by different rates of sequence divergence between homologous genes or by different rates of HGT of non-homologous genes. To investigate this, all genomic similarity comparisons involving dsDNA phages positioned within

intra-cluster boundaries were categorized as “temperate HGCF”, “temperate LGCF”, and “lytic”, based on the predicted lifestyle of the two phages and the position of the data on the plot (see Materials and Methods). For each pairwise comparison, genes were defined as “shared” or “unshared”, depending on whether they belong to a pham that is also present in the other genome, and several analyses were performed using either the whole genome data or the shared and unshared gene data subsets (Figure 2-9, Supplementary Table 2-2). Average genome size disparities are larger in temperate HGCF comparisons than in temperate LGCF or lytic comparisons (Figure 2-9A). Nucleotide distances between unshared genes in temperate HGCF comparisons are comparable to those in temperate LGCF and lytic comparisons (Figure 2-9B). In contrast, nucleotide distances of all shared genes in temperate HGCF comparisons are smaller than in temperate LGCF or lytic comparisons, when compared to whole genome nucleotide distance or to whole genome GCD (Figure 2-9B, C). The proportion of nucleotide sequence coding for all unshared genes relative to the nucleotide sequence coding for all genes in each comparison is correlated with changes in GCD regardless of lifestyle or evolutionary mode, suggesting that temperate HGCF comparisons are not caused by the gain and loss of a large number of tiny genes that skew GCD (Figure 2-9D). Similarly, the average size of unshared genes per comparison is comparable across lifestyle and evolutionary mode, and they are smaller than the average size of shared genes, as has been previously reported (Figure 2-9E)(Hatfull et al., 2010). These data indicate that evolutionary modes are associated with increased rates of horizontal gene transfer instead of increased rates of divergence between homologous sequences.

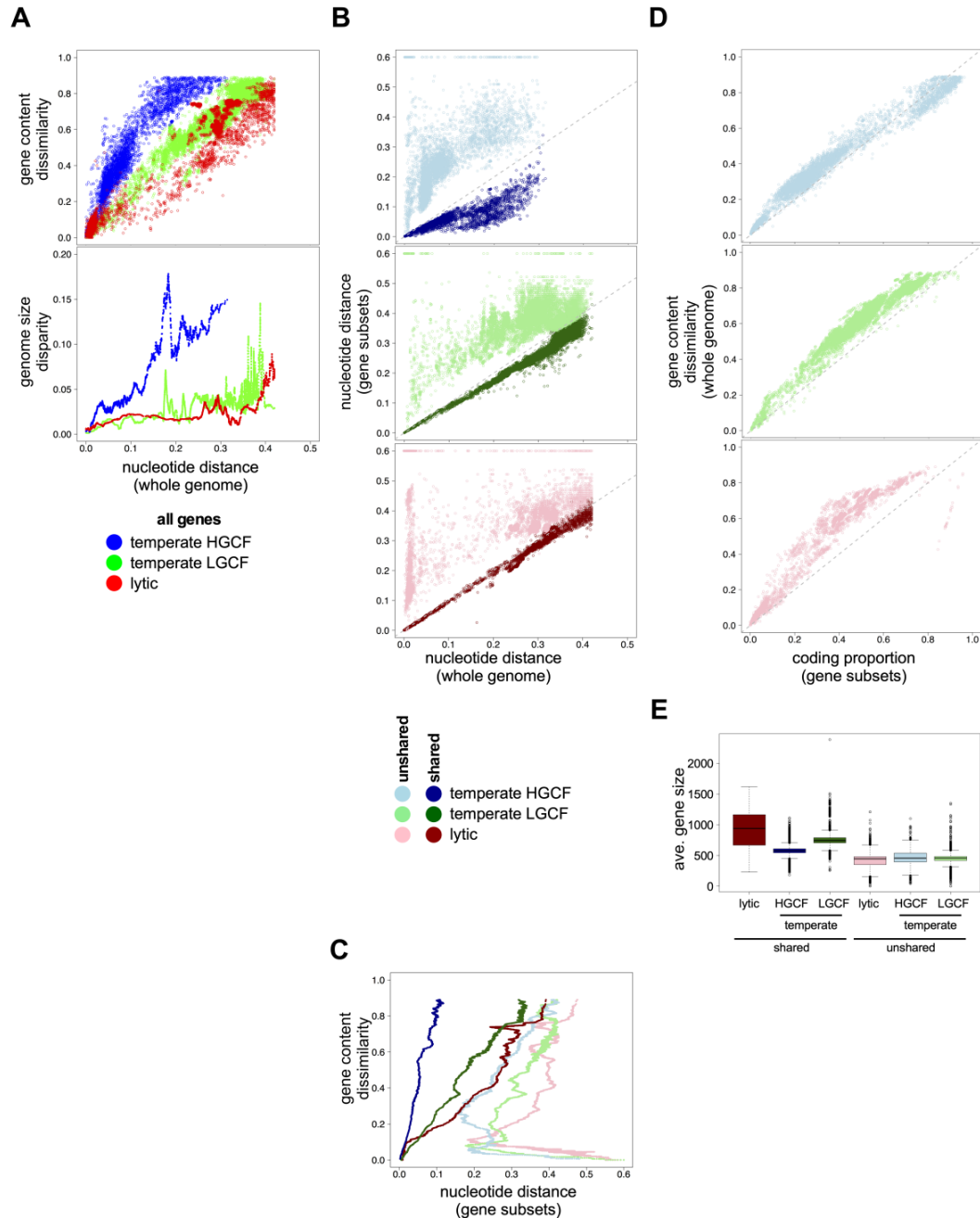


Figure 2-9. Evolutionary modes reflect different degrees of gene conservation.

(A) Genomic similarity scatter plot using a subset of comparisons from Figure 2-3A positioned within the intra-cluster sector (representing the most informative comparisons) and in which both lifestyles are known. (top) The comparisons are divided into three groups: temperate HGCF (blue), temperate LGCF (green), and lytic (red). (bottom) Sliding window averages of genome size disparities across whole genome nucleotide distances of the comparison subsets based on comparison type. **(B)** Nucleotide distances between only the shared (dark shade) or unshared (light shade) genes of each genome are plotted across whole genome nucleotide distances for the temperate

HGCF (top, blue), temperate LGCF (middle, green), and lytic (bottom, red) groups from panel A. The line at $y = x$ is plotted for reference. **(C)** Sliding window averages of shared and unshared gene distances (from panel B) across whole genome gene content dissimilarities (same colors as in panel B). **(D)** The proportion of coding sequence derived from unshared genes (relative to total shared and unshared gene coding sequence) is compared to gene content dissimilarity for temperate HGCF (top), temperate LGCF (middle), and lytic (bottom) groups. The line at $y = x$ is plotted for reference. **(E)** Average gene sizes for shared and unshared genes are compared between groups using box-and-whisker plots, in which the box represents the central 50% of the data, the black bar represents the median, and the points beyond the whiskers represent outliers. Figure adapted from (Mavrich and Hatfull, 2017).

2.3.6 Cluster A phages exhibit two evolutionary modes

Although most actinobacteriophage clusters exhibit a single evolutionary mode, phages in Cluster A exhibit two distinct modes (Figures 2-6D, 2-10A). The majority of Cluster A phages are categorized as LGCF, but some are categorized as HGCF (Figure 2-6D, Supplementary Table 2-1). In this dataset, there are over 200 phages grouped into Cluster A. They exhibit similar genomic architectures and gene regulatory strategies, but they are genetically diverse such that they have been subdivided into 17 subclusters (Mediavilla et al., 2000; Pope et al., 2011b). When genomic similarities between Cluster A phages are further evaluated based on subcluster division, it is clear that the two distributions are formed from two populations of Cluster A phages. All comparisons involving two Subcluster A1 phages distribute into the HGCF sector, and all comparisons involving two Cluster A phages that are not in Subcluster A1 distribute into the LGCF sector (Figure 2-10A). Subcluster A1 phages are distantly related to all other Cluster A phages, exhibited by the distribution of A1/non-A1 comparisons near the top right corner of the inter-subcluster sector (Figure 2-10A).

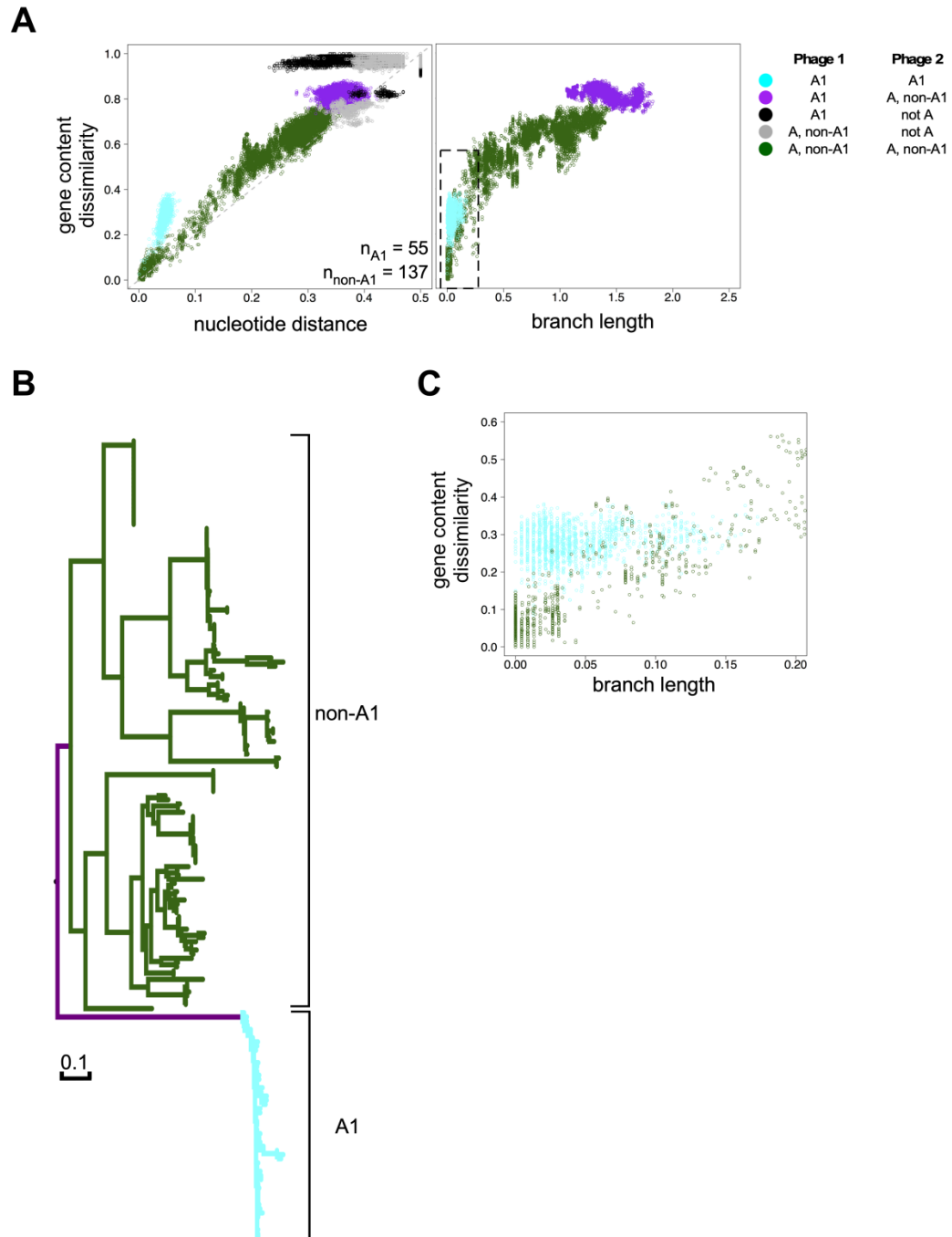


Figure 2-10. Cluster A phages exhibit two evolutionary modes.

(A)(left) Genomic similarity scatter plot involving *Mycobacterium* phage Cluster A-specific comparisons as in Figure 2-3A, with subcluster relationships highlighted. Cluster A comparisons involving two Subcluster A1 phages (cyan), two non-Subcluster A1 phages (dark green), one Subcluster A1 and one non-Subcluster A1 phage (purple), one Subcluster A1 and one non-Cluster A phage (black), and one non-Subcluster A1 and one non-Cluster A phage (grey). n = number of Cluster A phages in Subcluster A1 (A1) and Subclusters A2-A17 (non-A1). (right)

The same Cluster A gene content dissimilarities plotted against pairwise branch lengths from the phylogenetic tree in panel B. Box indicates the area of the plot enlarged in panel C. **(B)** Phylogenetic tree of all Cluster A phages based on structural/assembly genes (see Materials and Methods). All branches are colored as in panel A. **(C)** Enlarged area of right plot in panel A. Figure adapted from (Mavrich and Hatfull, 2017).

Several genes are completely conserved among all Cluster A phages, despite their genetic diversity. A phylogeny of Cluster A phages constructed from alignment of several of these conserved genes highlights that Subcluster A1 phages form a monophyletic clade separate from the other Cluster A phages, and they exhibit markedly shorter branch lengths (Figure 2-10B)(see Materials and Methods). When genomic similarity plots are constructed for Cluster A phages using phylogenetic branch lengths, which are corrected for evolutionary time, instead of Mash nucleotide distances, intra-Subcluster A1 comparisons continue to exhibit a distinct distribution from comparisons involving non-A1 phages (Figure 2-10A, C).

Similar results are observed when the phylogenetic analysis is extended to actinobacteriophages from other clusters representing different lifestyles and evolutionary modes. Separate phylogenies were constructed for phages in Cluster F (temperate HGCF), Clusters BD and K (temperate LGCF), and Cluster B (lytic) using alignments of several genes highly conserved within each cluster (see Materials and Methods). Using phylogenetic branch lengths for this larger dataset, temperate HGCF comparisons remain distinct from temperate LGCF and lytic comparisons in several ways. Temperate HGCF comparisons exhibit greater gene content dissimilarity and genome size disparities (Figure 2-11A, B). Even though the sizes of unshared genes are not substantially different, they have greater amounts of unshared genes categorized as orphans, and a greater degree of their coding sequence is derived from unshared genes (Figure 2-11C). Additionally, phams representing unshared genes are more widely distributed across actinobacteriophage clusters (Figure 2-11C). The data are consistent with

temperate HGCF phages exhibiting greater levels of HGT than temperate LGCF and lytic phages.

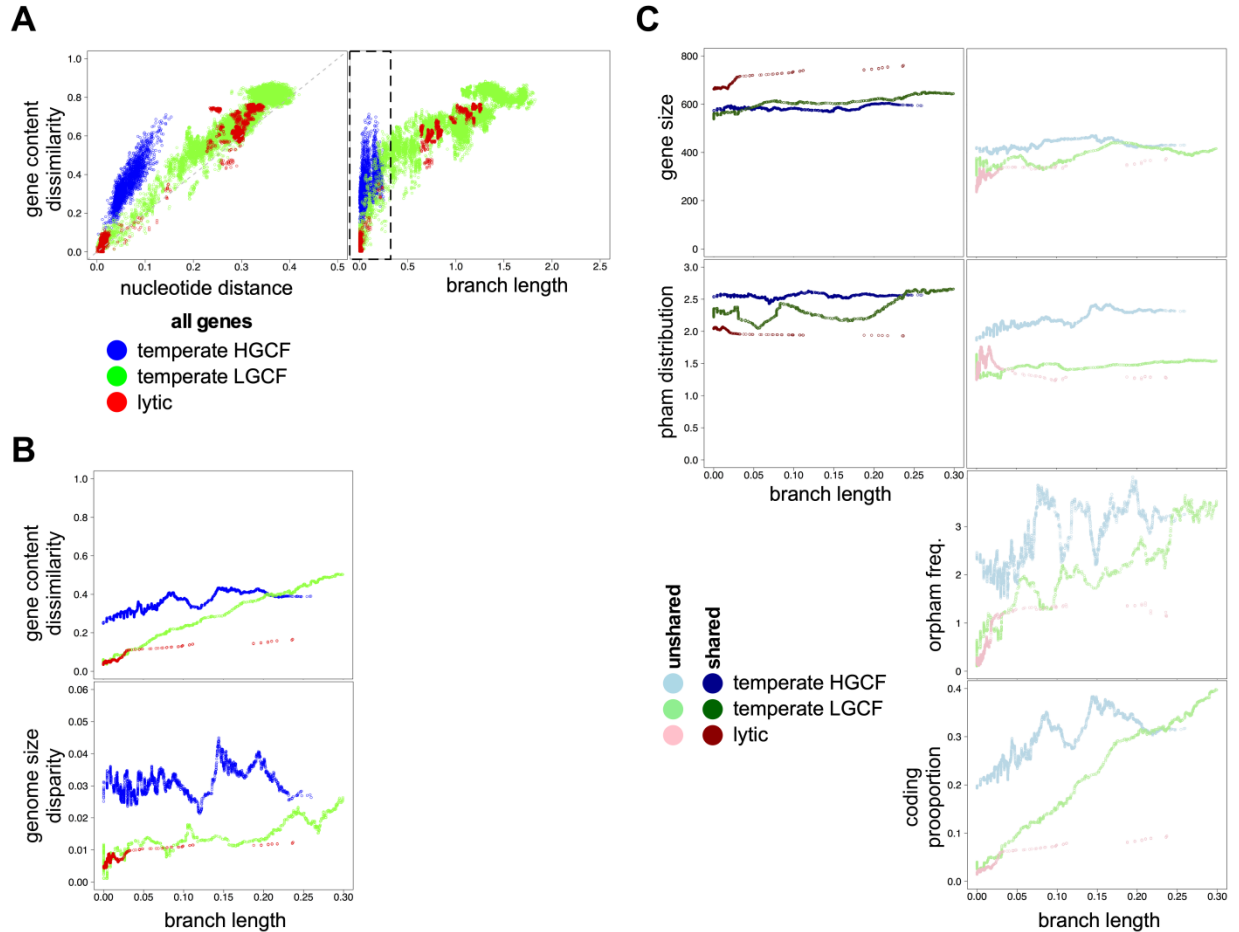


Figure 2-11. HGCF evolutionary mode is distinct in several genomic aspects.

(A) Genomic similarities are plotted for specific clusters in each phage group [Cluster F and Subcluster A1 (temperate HGCF, blue); Clusters BD, K, and A (non-A1)(temperate LGCF, green); and Cluster B (lytic, red)] using (left) Mash-based nucleotide distance and (right) branch lengths from cluster-specific alignment-based phylogenies, as in Figure 2-10A. Box indicates the area of the plot enlarged in panel B. (B) Sliding window averages of all intra-cluster (top) gene content dissimilarities and (bottom) genome size disparities for each group are plotted against branch lengths. (C) Sliding window averages of several metrics are plotted for (left) shared and (right) unshared genes in each comparison subset from panel A: gene size; pham distribution, orpham abundance, and coding sequence proportion (see Materials and Methods). Figure adapted from (Mavrich and Hatfull, 2017).

2.3.7 Temperate HGCF phages exhibit greater rates of HGT

Rates of HGT were measured by assessing the gain and loss of phams within a phylogenetic context, similar to studies of bacterial evolution (Puigbo et al., 2014). In the phylogeny specific to each cluster, the presence and absence of every pham in the cluster across the extant taxa was used by Count to identify the ancestral branches in which the pham was most likely to have been gained or lost (see Materials and Methods). Within the Cluster A phylogeny, Subcluster A1 phages exhibit nearly 10 times more pham gain and loss events than non-Subcluster A1 phages (Figure 2-12A). Very similar results are observed with phages in Cluster F compared to phages in Clusters BD, K, and B (Figure 2-12B). The phylogeny constructed from LysB proteins (constituting pham 21902) in Clusters A and F phages is consistent with these results (Figures 2-1B, 2-12C, D). The phylogenetic clade of Cluster A phages exhibits longer branch lengths than the clade of Cluster F phages, and Count predicts eight HGT events within Cluster A phages and none in Cluster F phages.

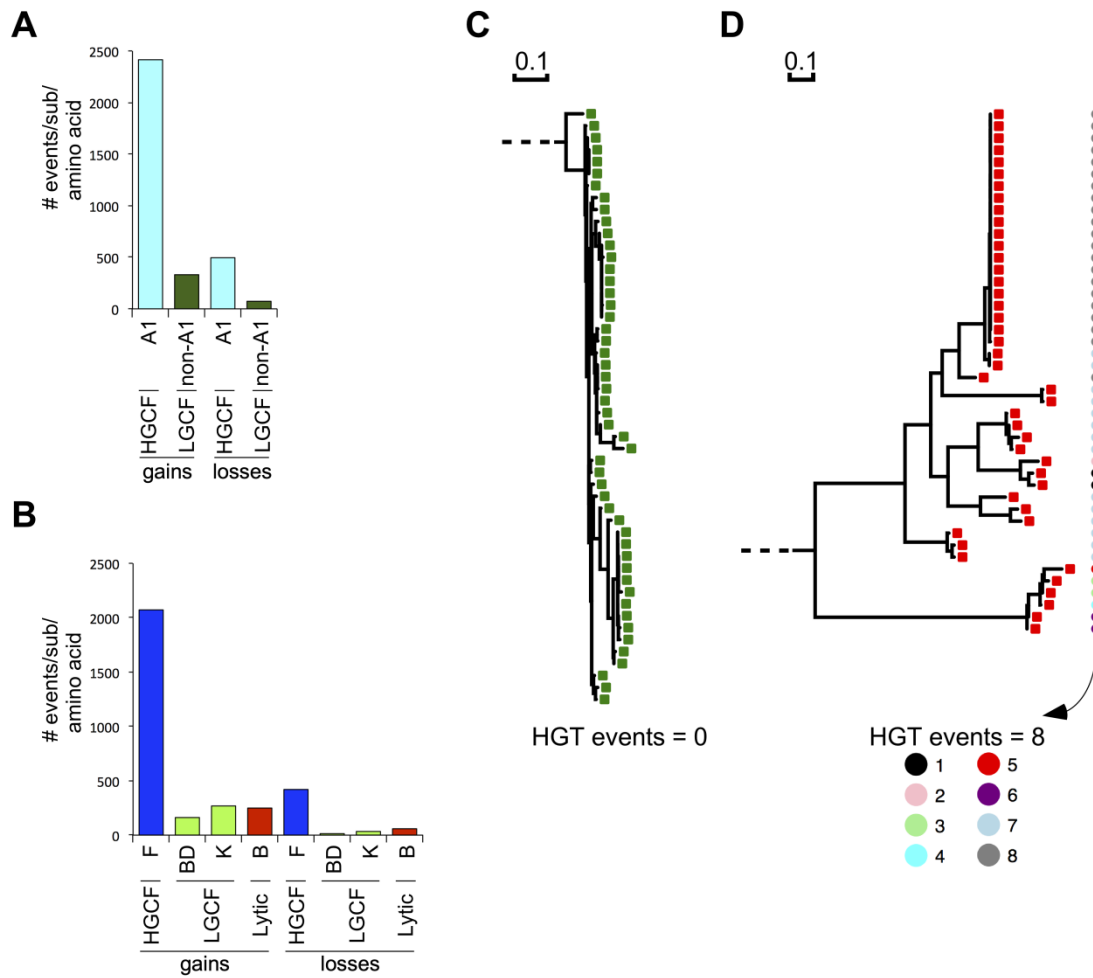


Figure 2-12. Evolutionary modes correlate with different rates of horizontal gene transfer.

(A) Bar graph of the predicted number of pham gains and losses per substitution per amino acid site for A1 and non-A1 phages (colored as in Figure 2-10A). (B) Bar graph of the predicted number of pham gains and losses as in panel A for additional representative clusters (colored as in Figure 2-11A). (C-D) LysB amino acid alignment-based phylogenetic subtrees from Figure 2-1B, with the number of HGT events predicted by Count indicated below, and taxa related to each HGT event labeled at right (colored circles). Figure adapted from (Mavrich and Hatfull, 2017).

The genetic basis for different rates of HGT is not obvious. Class 1 temperate phages do not exhibit substantial differences in GC% content, genome size, number of total genes, or the number of genes associated with specific stages of growth or functions (Figure 2-13A). The absolute or normalized rates of HGT for subsets of genes representing specific functions may vary between clusters, but they are not substantially different for Class 1 temperate phages compared to Class 2 temperate phages or lytic phages (Figure 2-13B, C). This is emphasized by Cluster A phages. All Cluster A phages are temperate or recent derivatives of temperate phages (Pope et al., 2015; Pope et al., 2011b). They exhibit similar genomic architecture and immunity systems to regulate lysogeny, and the same head packaging strategy and tail type (Pope et al., 2011b). Therefore, these shared characteristics among Cluster A phages highlight factors that are not likely to be determinants of evolutionary modes.

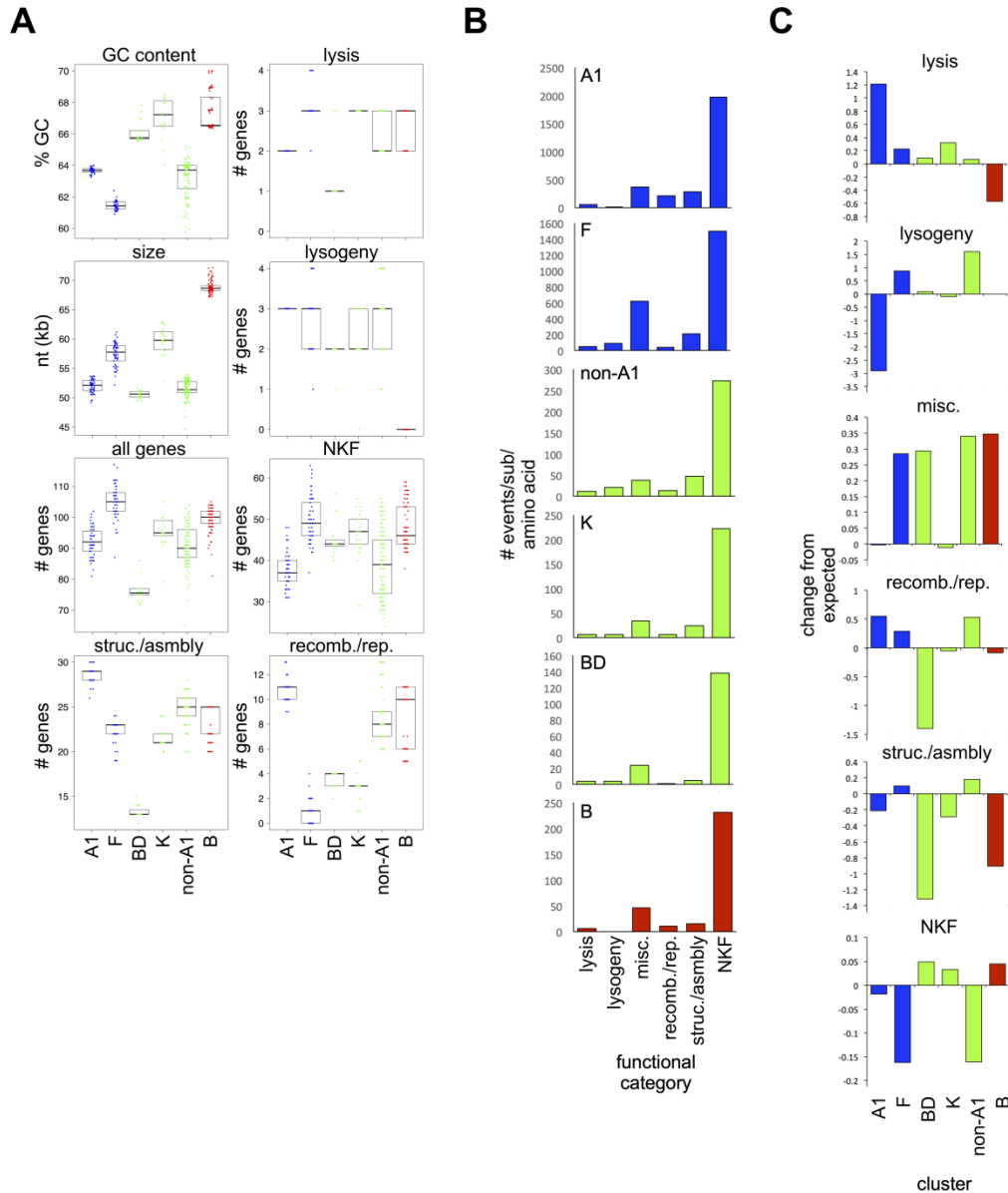


Figure 2-13. Evaluation of genome characteristics between evolutionary modes.

(A) Box plots comparing phages in specific representative clusters in the phylogenetic analysis, using several genome metrics such as GC% content, genome size (size), and the number of all genes or subsets of genes in functional categories per genome (struc/asmbly = structural and assembly; recomb/rep = recombination and replication; NKF = no known function). Each data point is a phage genome, and box plots depict the middle 50% of the data surrounding the median (black bar). **(B)** Bar plots of the absolute rates of HGT (gain and loss events combined) from Figure 2-12 based on gene functional categories for each group of phages. **(C)** Bar plots of normalized HGT rates for each functional category from panel B based on the proportion of phams associated with each category in the group of phages. For panels B-C colors are same as in Figure 2-12. Figure adapted from (Mavrich and Hatfull, 2017).

2.3.8 Evolutionary modes differ by host phyla

Phages may exhibit different evolutionary patterns based on their hosts, so genomic relationships were examined for phages of the five most predominant host phyla in this dataset (Figure 2-14, Supplementary Table 2-1). Similar to phages infecting hosts of Actinobacteria, phages infecting hosts of Proteobacteria exhibit two evolutionary modes (Figure 2-14A, E). The majority of lytic phages are categorized in the LGCF mode, and temperate phages exhibit both evolutionary modes. Phages infecting hosts of Firmicutes are distributed slightly differently: the majority of temperate phages are categorized as Class 1, and the distribution of LGCF comparisons are not continuous (Figure 2-14D). In this database, there are few phages infecting hosts of Bacteroidetes and since they are too genetically diverse, the evolutionary mode for most of them cannot be determined, but in general their genomic comparisons are predominantly distributed across the HGCF sector (Figure 2-14B). Phages infecting hosts of Cyanobacteria exhibit the most marked differences in evolutionary patterns (Figure 2-14C). Although there are fewer than 100 of these phages, they exhibit a sufficient genetic spectrum to determine evolutionary mode, and all phages are categorized as LGCF regardless of lifestyle.

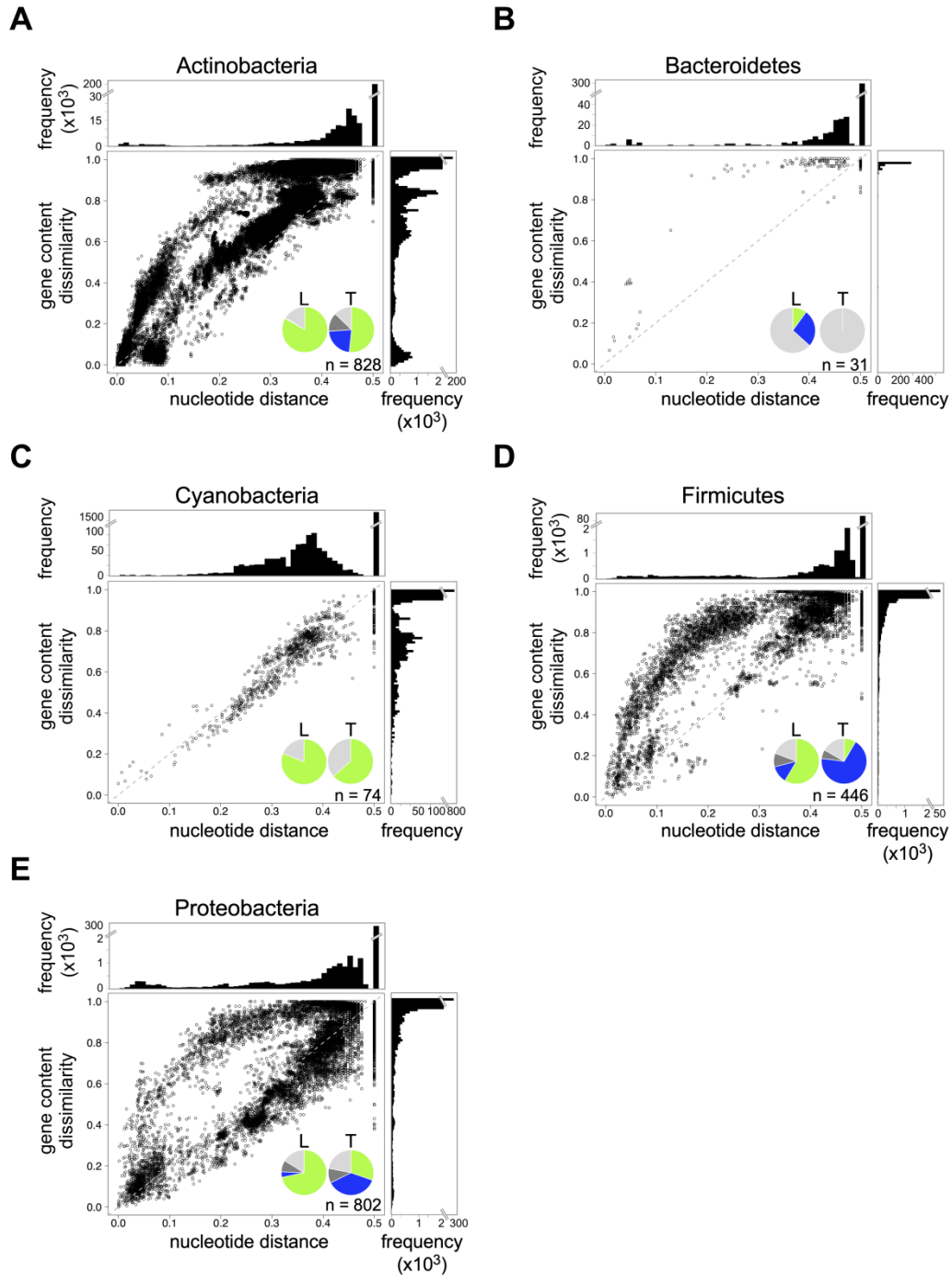


Figure 2-14. Host phyla exhibit diversity in evolutionary modes.

(A)-(E) Genomic similarity scatter plots as in Figure 2-3A, using subsets of data based on the five most predominant host phyla. Pie charts reflect the proportion of phages of each host phylum that are predicted to be in each mode for each predicted lifestyle, as in Figure 2-7E (L = lytic, T = temperate). n = number of phages present in the host phylum that were used for the analysis. Figure adapted from (Mavrich and Hatfull, 2017).

2.3.9 Implications of evolutionary modes

Although the genetic basis of phage evolutionary modes is not obvious, Class 1 temperate phages may have a greater impact on their environment than Class 2 temperate phages in several ways. In general, temperate and lytic phages are not closely related and exhibit different gene content, but genomic similarity comparisons between temperate and lytic phages are distributed along the HGCF mode (Figure 2-7A). Similarly, genomic comparisons involving phages with different tail types are distributed along the HGCF mode (Figure 2-8B). Genomic relationships involving λ are distributed along the HGCF mode (Figure 2-15A). Genomic relationships involving phages that encode virulence factors associated with bacterial pathogenicity are distributed along the HGCF mode and the majority of them are classified as Class 1 temperate phages (Figure 2-15B, Supplementary Table 2-1). Lastly, genomic relationships involving Cluster N temperate phages, which encode diverse phage defense systems (Dedrick et al., 2017a), are distributed along the HGCF mode (Figure 2-15C, Supplementary Table 2-1). Therefore, many characteristics commonly associated with temperate phages may be more closely associated with Class 1, instead of Class 2, temperate phages.

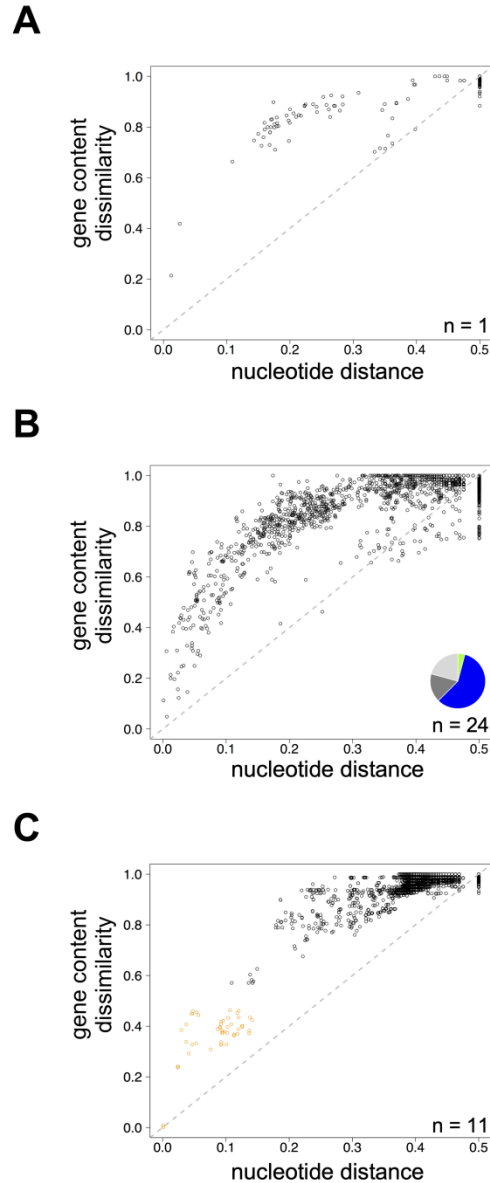


Figure 2-15. Many well-studied temperate phages are associated with HGCF.

Genomic similarity scatter plots as in Figure 2-3A using **(A)** all comparisons involving the enterobacteria phage λ and **(B)** all comparisons involving at least one phage previously characterized as encoding virulence factors. Pie chart reflects the proportion of these phages predicted to be in each mode, as in Figure 2-7E. n = number of phages encoding virulence factors used for the analysis. **(C)** Genomic similarity scatter plot as in Figure 2-5 involving all Cluster N intra-cluster (orange) and inter-cluster (black) comparisons. n = number of Cluster N phages used for the analysis. Figure adapted from (Mavrich and Hatfull, 2017).

2.3.10 Quantification of phage genetic isolation using MaxGCDGap

Genetic diversity among isolated phages is heterogeneous (Hatfull, 2010). Even after sequencing thousands of actinobacteriophages, some clusters remain much larger than others (such as the Cluster A phages, representing 25-30% of all *Mycobacterium* phages), while some phages exhibit no close genetic relatives and remain as singletons. Genetic diversity within clusters is also heterogeneous. For example, the 192 Cluster A phages and 59 Cluster F phages exhibit a nearly complete spectrum of genetic relationships, spanning nucleotide distances from 0 to 0.5 and gene content dissimilarities from 0 to 1 (Figures 2-6C, 2-10A). In contrast, the 39 phages in Cluster BU are very similar to each other but exhibit no close relationships to any other non-BU phage, reflected by the large gap between the distribution of their intra-cluster and inter-cluster comparisons (Figure 2-6C). Although the genomic similarity plot reflects this heterogeneity, it is difficult to directly compare the degree of genetic isolation associated with each phage or group of phages.

Identifying the largest gap in a distribution of genomic relationships across either the GCD or nucleotide distance axes, or both, can provide a quantified metric of genetic isolation. For example, when all GCDs involving the *Gordonia* Singleton phage GMA2 with any other actinobacteriophage are ranked by magnitude, it is apparent that GMA2 is genetically distant from all other actinobacteriophages (Figure 2-16A). Consequently, the maximum gap in the ranked GCDs (MaxGCDGap) is ~ 1 (approaching the largest possible GCD gap size). In contrast, since the *Gordonia* Cluster CS phage Monty exhibits shared gene content with several phages, there are several GCDs less than 1. Therefore, Monty has a MaxGCDGap of only ~ 0.3 , occurring between the comparison involving Cluster CS *Gordonia* phage Kvothe and the comparison involving *Rhodococcus* Singleton phage ReqiDocB7 (Figure 2-16A).

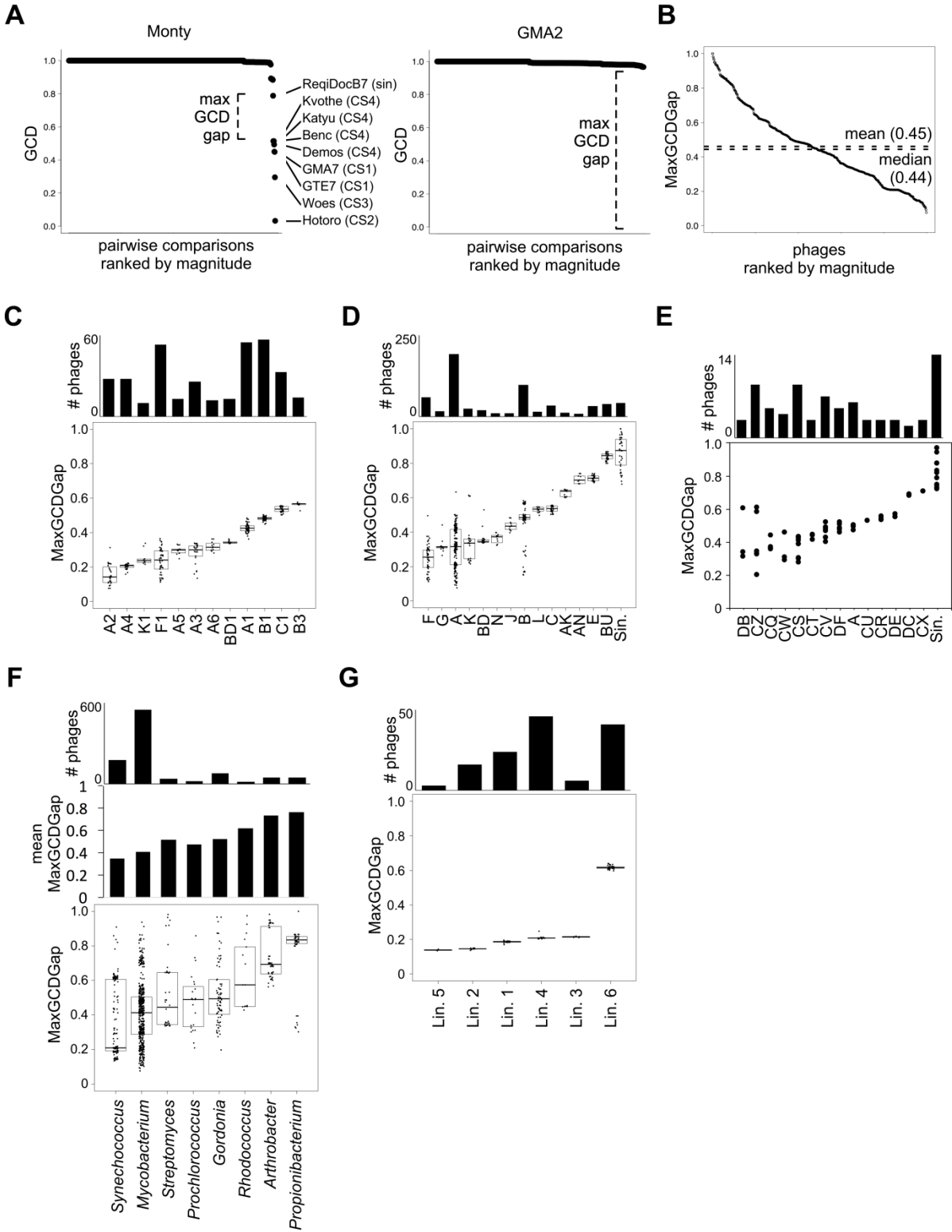


Figure 2-16. Measuring degrees of phage genetic isolation using MaxGCDGap.

(A) GCDs between representative *Gordonia* phage (left) Monty or (right) GMA2 and all other phages in the database, ranked by magnitude. Phages involved in comparisons with a GCD of < 0.8 are highlighted. The maximum gap in GCD values (MaxGCDGap) was identified for each phage. For example, Cluster CS Monty's MaxGCDGap occurs between *Gordonia* phage Kvothe (Cluster CS) and the Singleton (Sin.) *Rhodococcus* phage ReqiDocB7. Singleton phage GMA2 has no close relatives, and the MaxGCDGap is large and approaches 1.0. (B) All phage-specific MaxGCDGap values ranked by magnitude, with the mean and median indicated. Each data point represents a single phage genome. (C-D) Box plot distribution of actinobacteriophage-specific MaxGCDGaps from panel B grouped by (C) subcluster or (D) cluster and ranked by median. Boxes reflect the central 50% of the data, with the median as a black bar, and the individual MaxGCDGap values are superimposed. Only the most abundant subclusters or clusters are plotted. (E) All *Gordonia* phage-specific MaxGCDGaps grouped by cluster and ranked by median. Each data point represents a single phage genome. (F) Box plot distribution of phage-specific MaxGCDGaps as in panels C and D but grouped by host genus and with mean MaxGCDGaps displayed above. Only the most abundant genera are plotted. (G) Box plot distribution of *Synechococcus* phage-specific MaxGCDGaps as in panels C, D, and F, grouped by the six previously identified lineages (Lin.) (Gregory et al., 2016). In all panels, the topmost bar chart indicates the number of phages per group. Figure adapted from (Pope et al., 2017).

When the MaxGCDGap is computed for all phages and ranked by magnitude, it is clear that phages in this database reflect a spectrum of genetic isolation (Figure 2-16B, Supplementary Table 2-3). Some phages exhibit a MaxGCDGap approaching 1, indicating that none of their genes are shared with any other phages in the database. Some phages exhibit a MaxGCDGap approaching 0, reflecting that they exhibit a complete genetic spectrum of relationships with other phages in the database.

Phage genetic isolation varies between subclusters, clusters, and host. Phages in some subclusters, such as B3 and C1, exhibit MaxGCDGaps $\sim 0.5-0.6$, while phages in other subclusters, such as A2 and A4, exhibit MaxGCDGaps $\sim 0.1-0.2$ (Figure 2-16C). Thus, although there is sufficient genetic diversity within Clusters B and C to warrant subcluster divisions, phages in those subclusters are nevertheless relatively isolated from other phages in the database compared to phages in other subclusters, such as A2 and A4. Similarly, phages in some clusters

such as F, G, and A are not very isolated, but phages in clusters E and BU are nearly as isolated from phages outside of their cluster as many singletons are (Figure 2-16D). Even among phages that infect hosts of the same genus, such as *Gordonia*, there is nevertheless a large degree of variation in phage genetic isolation (Figure 2-16E). Of the 79 *Gordonia* phages, 65 have been grouped into 14 clusters, and 14 remain as singletons. Singleton phages exhibit large MaxGCDGaps, as expected. Phages of different clusters exhibit a range of MaxGCDGaps; phages in Cluster CX have MaxGCDGaps nearly as large as some of the singleton phages, and phages in Clusters DB and CZ have some of the smallest MaxGCDGaps. MaxGCDGaps are not correlated with the number of phages present in each subcluster or cluster (Figure 2-16C, D, E). Additionally, phages infecting specific hosts may be more genetically isolated than phages in other hosts. Phages infecting *Propionibacterium* and *Arthrobacter* genera are relatively isolated, even though similar amounts of phages have been sequenced for these hosts as for less isolated phages infecting *Streptomyces* (Figure 2-16F).

Recent investigations have reported that marine phages infecting cyanobacterial hosts, including *Synechococcus* and *Prochlorococcus* genera, do not exhibit the same degree of genetic mosaicism as reported for actinobacterial hosts, and several distinct cyanophage lineages can be observed (Gregory et al., 2016). Genomic similarity plots indicate that cyanophages evolve only within the LGCF mode (Figure 2-14C). To further evaluate cyanophage evolution, I compared the degree of their genetic isolation to previously reported lineages using over 200 cyanophages, many of which were also used for the evolutionary mode analysis (Supplementary Table 2-3, see Materials and Methods). Phages infecting cyanobacterial hosts exhibit wide ranges of MaxGCDGaps (Figure 2-16F). Additionally, phages infecting *Synechococcus* hosts exhibit the smallest median MaxGCDGap when compared to phages of other hosts, and this does not

correlate with the number of phages isolated per host (Figure 2-16F). These results suggest cyanophages are no less genetically isolated than actinobacteriophages (Figure 2-16F). Furthermore, most MaxGCDGaps within previously reported cyanobacteriophage lineages (Gregory et al., 2016) are less than ~ 0.2 , indicating they exhibit a smaller degree of isolation than many actinobacteriophage subclusters (Figure 2-16C, G). Thus, conflicting reports regarding phage diversity and mosaicism may be due to utilization of analyses that rely on different evolutionary timescales, instead of due to phages that exhibit different evolutionary pressures.

2.4 DISCUSSION

The biological basis of different evolutionary modes is not obvious. However, since individual phages, and groups of genetically related phages predominantly exhibit trajectories through only one mode, the factors constraining them to either mode are likely to be extended over long evolutionary timeframes. Since subsets of Cluster A *Mycobacterium* phages exhibit both evolutionary modes, the modes are not dependent on overall genome architecture or replication strategy. The two modes may be caused by the length of time phages reside within the host cell. The time that temperate phages spend as resident prophages has been postulated to play a factor in their increased genetic diversity compared to lytic phages (Chopin et al., 2001). This may only be the case for Class 1 temperate phages though, and Class 2 temperate phages may reside within the host for similar lengths of time as lytic phages. We do not fully understand how long temperate phages reside in the host between periods of lytic growth, let alone the factors that impact those timeframes. Lambdoid prophages monitor the host intracellular

environment and can enter lytic growth under different conditions (Rozanov et al., 1998). In this case, the two evolutionary modes may reflect extended prophage states due to more stable conditions of the host cell, and by extension the broader environmental context. Alternatively, the two modes could be driven by genetic factors such as the presence or absence of specific genes such as recombinases that facilitate gene exchange, or the location on the host chromosome where integration occurs (Bobay et al., 2013). In this scenario, the two modes reflect increased recombination rates. No specific genes associated with the HGCF mode have been identified yet, though. Lastly, the two modes may reflect different gene pools that phages have access to (Hendrix et al., 1999; Jacobs-Sera et al., 2012). Horizontal gene transfer events can involve two homologous genes, and changes in gene content, measured using phams, do not reflect these events. Phages in the LGCF mode may not have access to gene pools of the same size and diversity as phages in the HGCF mode, so even though they may encounter the same levels of HGT, it would not be reflected in gene content dissimilarity.

Several types of experiments may provide insight into what causes the evolutionary modes. Increased sequencing of phages derived from specific environments could reveal whether evolutionary modes are correlated with environmental factors. For instance, phages identified from metagenomics of the gut microbiome (Kim and Bae, 2018; Manrique et al., 2016) or the marine environment (Roux et al., 2015; Roux et al., 2014) may be associated more with one mode than the other. Similarly, prophages are commonly identified in bacterial genomes (Bobay et al., 2013; Casjens, 2003; Lawrence et al., 2001; Touchon et al., 2016). Since these prophages are directly associated with their hosts, an analysis of the evolutionary modes of these prophages could determine whether they are associated with one mode more than the other. Using these larger datasets that reflect a direct linkage between the prophage and host, more advanced

computational tools, such as machine learning, could be implemented to identify what genetic factors are associated with one evolutionary mode compared to the other. The Cluster A phages can serve as a good control, since they exhibit both evolutionary modes.

Measuring genomic similarities and degrees of genetic isolation can resolve apparent disagreements in patterns of phage evolution. In contrast to evolutionary studies involving phages that infect hosts of Proteobacteria (Grose and Casjens, 2014) and Actinobacteria (Pope et al., 2015), studies involving phages infecting hosts in the phylum Cyanobacteria have suggested that cyanophage populations are genetically discrete (Deng et al., 2014; Gregory et al., 2016), and that the phage gene pool is not as large as previously estimated (Ignacio-Espinoza et al., 2013). However, the degree to which marine phages are temperate is not well understood (Mann, 2003; Paul et al., 2002), 85% of cyanophages in this dataset are predicted to be lytic, and the genomic similarity plot reveals that they evolve within the LGCF mode. Additionally, the MaxGCDGap analysis indicates that cyanophage lineages are no more genetically isolated than actinobacteriophage subclusters are. Therefore, conflicting conclusions regarding phage evolutionary patterns may arise if evaluations are performed using groups of phages that span different evolutionary timescales and represent different evolutionary modes.

The variation in evolutionary modes observed between different host phyla may reflect different host-related selective pressures that shape phage diversity. Alternatively, they may be an artifact of biased phage isolation techniques that recover particular subsets of phages. Temperate cyanophages may be abundant, and phages infecting Firmicutes that evolve within the LGCF mode may also be abundant, but they may not be readily recovered unless new techniques are developed. Additionally, phage diversity within each host phyla may be impacted by host diversity. The lab-derived host strains that are used to isolate phages are likely different

from the host strains the phages have predominantly infected in the environment. Isolating phages using different types of hosts may impact the distribution of phages in each evolutionary mode in each host phylum.

The systematic grouping of phages into clusters and subclusters reflects varying degrees of genomic relationships and facilitates downstream analyses. However, it is a manual, time-intensive process and is routinely applied primarily to actinobacteriophages [although the strategy has been adopted for classification of phages infecting enterobacteria (Grose and Casjens, 2014)]. Additionally, it does not fully capture reticulate relationships that extend between these delineations. The bivariate genomic similarity plot highlights the complexity of trying to group phages: the use of a single metric, such as shared gene content, presence/absence of gene signatures, or divergence of structural genes, do not account for different evolutionary modes. The calibration of GCD and nucleotide distance metrics using actinobacteriophage clusters and subclusters can enable the genomic similarity plot to be used as a bivariate tool to rapidly and automatically group phages at multiple levels of genomic relationships. Phages of other host phyla can be grouped at cluster-level and subcluster-level divisions, and all phages can be grouped at customized degrees of similarity with finer resolution (such as at species-level nucleotide distances or by evolutionary mode) to address specific research questions.

3.0 CHARACTERIZATION AND INDUCTION OF BIFIDOPROPHAGES

The data in this chapter was published in the journal *Scientific Reports* (Mavrich et al., 2018), and figures have been adapted for this work. Although I performed the majority of the experiments and analyses, some experiments were performed by others. Joana Oliveira prepared flow cytometry *Lactococcus* phage control samples. Charles M.A.P. Franz and Horst Neve performed electron microscopy of induced lysates. Gabriel Andrea Lugli and Marco Ventura sequenced the induced lysates and created the phylogenetic tree of bifidophages genomes. Francesca Bottacini performed the analysis of bifidobacterial whole genome sequencing reads to assess *Rin* shufflon inversion patterns.

3.1 INTRODUCTION

By impacting the growth and evolution of their bacterial hosts, bacteriophages play powerful roles in their environment (Brussow, 2001; Suttle, 2005). One environment that is not well understood is the human gut microbiome. This dynamic microbial community is comprised of hundreds of species spanning numerous phyla and genera (Qin et al., 2010), and their complex interactions may impact human health (Shreiner et al., 2015). Phages are abundant in this environment (Dutilh et al., 2014; Manrique et al., 2016; Stern et al., 2012) and can be used to artificially modulate the community (Reyes et al., 2013). Identifying novel phages of diverse hosts can enhance our understanding and control of this environment.

Bacteria of the genus *Bifidobacterium* are prevalent and important members of the gut (Arbolea et al., 2016). These organisms are Gram-positive and anaerobic, and they are members of the phylum Actinobacteria. They are the prominent bacterial genera populating the gut at birth, persist across the human lifespan, and are associated with a positive health status (Arbolea et al., 2016). Two of the most abundant bifidobacterial species in the infant gut are *B. breve* and *B. longum* (Turroni et al., 2012). However, isolation and propagation of bifidobacterial phages (bifidophages) infecting these species have not been reported.

Currently there are over 2,900 isolated and sequenced actinobacteriophages from over 14 host genera (<https://phagesdb.org>). The majority of phages in this collection infect a single host genus, *Mycobacterium*, and expanding the phylogenetic breadth of host genera, such as *Propionibacterium* (Marinelli et al., 2012), *Arthrobacter* (Klyczek et al., 2017), and *Gordonia* (Pope et al., 2017), has enabled comparative analyses that continually enhance our understanding of phage biology, diversity, evolution, and host interactions. Since phages have not yet been isolated for over 100 actinobacterial genera, our understanding of actinobacteriophage diversity is limited, and the lack of characterized bifidophages specifically limits our understanding of actinobacteriophage biology in the unique gut microbial environment.

Bifidobacterial comparative analyses suggest that bifidophages are abundant. The majority of sequenced bifidobacterial strains are predicted to contain complete or cryptic prophages (Lugli et al., 2016b; Ventura et al., 2005b), such as Binf-1 in *B. longum* subsp. *infantis* ATCC 15697 (Ventura et al., 2009). Genetically related prophages are present in multiple species, suggesting they have either broad or dynamic host range specificities. Many strains contain phage defense systems such as CRISPR arrays with spacers that match many of the predicted prophages (Briner et al., 2015; Ventura et al., 2009), suggesting frequent host-

phage interactions. In addition, excision of some prophages has been induced by treatment with chemicals such as hydrogen peroxide or mitomycin C (Milani et al., 2014). Mitomycin C induction of prophages from more distantly-related and uncharacterized *B. moukalabense* and *B. choerinum* species produce complete phage particles (Milani et al., 2014). There have been no reports of inducible phage particles from bifidobacterial strains that are more closely associated with the human gut.

Here, I identified and characterized three groups of prophages in several *B. breve* and *B. longum* strains. These prophages are integrated into a tmRNA gene, the tRNA^{Met} gene, or the *dnaJ₂* gene, a novel phage integration site that is unique to Actinobacteria. I successfully induced these prophages using mitomycin C and show that they replicate and form complete phage particles. Some of them contain a phase variation locus that modulates receptor binding protein (RBP) genes with a tyrosine invertase (Rin), analogous to other characterized enterobacteria phage phase variation loci, Min and Cin.

3.2 MATERIALS AND METHODS

3.2.1 Bacterial strains

Several bifidobacterial strains were used in this study, as indicated in Table 3-1 [adapted from (Mavrich et al., 2018)].

Table 3-1. Bifidobacterium genomes used in this study.

Strain	Accession	Reference	Strain type (for data analysis)	GC%
<i>B. choerinum</i> LMG 10510	JGYU000000000	Milani <i>et al.</i> , 2014	Lysogen	65.5
<i>B. moukalabense</i> DSM 27321	AZMV010000000	Lugli <i>et al.</i> , 2014	Lysogen	59.9
<i>B. longum infantis</i> ATCC 15697	AP010889	Fukuda <i>et al.</i> , 2011	Lysogen	59.9
<i>B. breve</i> JCM 7017	CP006712	Bottacini <i>et al.</i> , 2014	Non-lysogen	58.7
<i>B. breve</i> NCIMB 702258	CP006714	Bottacini <i>et al.</i> , 2014	Non-lysogen	58.7
<i>B. breve</i> UCC2003	CP000303	Motherway <i>et al.</i> , 2011	Non-lysogen	58.7
<i>B. breve</i> 082W4-8	CP021555	Bottacini <i>et al.</i> , 2017	Predicted lysogen	58.8
<i>B. breve</i> 180W8-3	CP021557	Bottacini <i>et al.</i> , 2017	Predicted lysogen	58.8
<i>B. breve</i> 139W4-23	CP021556	Bottacini <i>et al.</i> , 2017	Predicted lysogen	58.6
<i>B. breve</i> 017W4-39	CP021554	Bottacini <i>et al.</i> , 2017	Predicted lysogen	58.7
<i>B. breve</i> 215W4-47a	CP021558	Bottacini <i>et al.</i> , 2017	Predicted lysogen	59.3
<i>B. longum longum</i> CCUG 30698	CP011965	O'Callaghan <i>et al.</i> , 2015	Predicted lysogen	60.2
<i>B. longum longum</i> 157F	AP010890	Fukuda <i>et al.</i> , 2011	Predicted lysogen	60.1
<i>B. breve</i> JCM 1192	AP012324	<i>Not applicable</i>	Predicted lysogen	58.9
<i>B. breve</i> 689b	CP006715	Bottacini <i>et al.</i> , 2014	Predicted lysogen	58.7

3.2.2 Prophage characterization

Several bifidophages were analyzed in this study, as indicated in Table 3-2 [adapted from (Mavrich et al., 2018)]. Prophages present in *B. breve* 082W4-8, *B. breve* 180W8-3, *B. breve* 139W4-23, *B. breve* 017W4-39, and *B. breve* 215W4-47a strains were previously reported (Bottacini et al., 2017). Prophages integrated at the homologous locus in other *B. breve* and *B. longum* strains were identified by BLAST (Altschul et al., 1990) using the predicted integrases. Gene functions were predicted with BLAST (Altschul et al., 1990) and HHpred (Soding et al., 2005). ProgressiveMauve (Darling et al., 2010) whole genome alignment was used to precisely identify integration sites, prophage sizes, and attachment sites. Prophage sequences were extracted from the host genome, and their nucleotide sequence and gene content were compared using Gepard (Krumsiek et al., 2007) dot plot analysis and Phamerator (Cresawn et al., 2011). The phylogenetic analysis using whole genome alignments to compare newly identified bifidophages with previously reported bifidophages was performed by Gabriele Lugli as previously described (Lugli et al., 2016b).

Table 3-2. Bifidoprohage genomes analyzed in this study.

Prophage	Host strain	Reference	Left boundary ¹	Right Boundary ¹	Size (nt)	GC%	Integration Locus	Group ⁴
Bb48phi1	<i>B. breve</i> 082W4-8	Bottacini <i>et al.</i> , 2017	1,193,403	1,232,804	39,402	61.3	<i>dnaJ</i> ₂ (BB082W48_0987)	3
Bb83phi1	<i>B. breve</i> 180W8-3	Bottacini <i>et al.</i> , 2017	1,179,127	1,219,381	40,255	61.1	<i>dnaJ</i> ₂ (BB180W83_0986)	3
Bb423phi1	<i>B. breve</i> 139W4-23	Bottacini <i>et al.</i> , 2017	1,302,639	1,342,715	40,077	60.9	<i>dnaJ</i> ₂ (BB139W423_1102)	3
Bb439phi1	<i>B. breve</i> 017W4-39	Bottacini <i>et al.</i> , 2017	1,194,717	1,234,971	40,255	61.1	<i>dnaJ</i> ₂ (BB017W439_1000)	3
Binf-1 ²	<i>B. longum infantis</i> ATCC 15697	Ventura <i>et al.</i> , 2009	1,288,185	1,330,866	42,682	61.1	<i>dnaJ</i> ₂ (BLIJ_1123)	3
Bl30698phi1	<i>B. longum longum</i> CCUG 30698	<i>this study</i>	1,375,860	1,336,483	39,378	61.1	<i>dnaJ</i> ₂ (BBL306_1177)	3
Bl157phi1	<i>B. longum longum</i> 157F	<i>this study</i>	1,246,936	1,207,777	39,160	60.9	<i>dnaJ</i> ₂ (BLIF_1084)	3
Bb447phi1	<i>B. breve</i> 215W4-47a	Bottacini <i>et al.</i> , 2017	1,694,074	1,735,438	41,365	58.6	tmRNA ³	4
Bb1192phi1	<i>B. breve</i> JCM 1192	<i>this study</i>	1,468,985	1,509,872	40,888	59.4	tmRNA ³	4
Bb423phi2	<i>B. breve</i> 139W4-23	Bottacini <i>et al.</i> , 2017	420,545	438,765	18,221	61.4	tRNA ^{Met}	1
689b-1	<i>B. breve</i> 689b	Bottacini <i>et al.</i> , 2014	372,287	390,546	18,260	61.4	tRNA ^{Met}	1

¹Coordinates based on sequence orientation in published GenBank record. For prophages integrated at *dnaJ*₂ gene, left and right boundaries are based on empirically determined site of strand exchange. Otherwise, boundaries are determined based on entire attachment sites. ²Prophage coordinates and integration locus modified from previous description (Lugli *et al.*, 2016b; Ventura *et al.*, 2009) due to analysis with other prophages in this study. ³tmRNA is not annotated in this GenBank record. ⁴Group = as determined by whole genome phylogenetic analysis in Figure 3-3C.

3.2.3 Phamerator database construction

The database *Actinobacteriophage_1060* was created using Phamerator (Cresawn et al., 2011), consisting of 1,060 actinobacteriophages and prophages, and is available online (http://phamerator.webfactional.com/databases_Hatfull). Genes are grouped into related gene families (“phams”) using kClust (Hauser et al., 2013).

3.2.4 Optimization of mitomycin C induction

Mitomycin C concentration was optimized using 96-well microtiter plates. Wells with 500 μ l RCM were inoculated from *B. breve* JCM 7017, *B. breve* UCC2003, and *B. breve* 017W4-39 cultures, grown for 8 h, and treated with a 10-fold serial titration of mitomycin C (ranging from 0.0003 μ g/ml to 3 μ g/ml) for 14 h. Growth inhibition was observed for concentrations at and above 0.03 μ g/ml. Similar inhibitory profiles were observed with 0.03 μ g/ml mitomycin C for 2.5 ml, but not 50 ml, cultures. Therefore, for 50 ml cultures, 0.3 μ g/ml mitomycin C was used.

3.2.5 Bifidobacterial growth and mitomycin C induction

Bifidobacterial strains were grown in 10 ml Reinforced Clostridial Medium (RCM) in a conical tube inoculated directly from freezer stock and grown to saturation overnight at 37°C in an anaerobic chamber. For mitomycin C induction tests, 50 ml RCM was inoculated from saturated culture at an $OD_{600nm} \sim 0.05$, inverted several times for gentle mixing, and grown at

37°C in anaerobic chamber for 4-5 h without shaking. When the culture reached an OD_{600nm} of 0.15-0.25, mitomycin C was added to 0.3 µg/ml, inverted several times for gentle mixing, and incubated at 37°C in an anaerobic chamber for 15-20 h. Final OD_{600nm} was recorded and the entire culture was centrifuged in a table-top centrifuge with swinging bucket rotor at 9,148 x g for 20 min with slow deceleration. The supernatant was transferred to a 50 ml syringe, filtered using a 0.45 µm filter, and stored at 4°C. Each sample was paired with an untreated control in which the 50 ml culture was allowed to grow to saturation in the absence of mitomycin C.

3.2.6 Induction verification using PCR

Prophage induction was confirmed by PCR amplification across the *attP* site that forms after excision and circularization. One pair of primers was designed to test *dnaJ2*-integrated phages (oTM202 and oTM203)(Appendix B). Separate primer pairs were designed for Bb447phi1 (oTM206 and oTM207) and Bb423phi2 (oTM208 and oTM209). Induction of previously described prophages in strains *B. choerinum* LMG 10510 and *B. moukalabense* DSM 27321 was confirmed using previously described primers (Lugli et al., 2016b). Amplification proceeded in 25 µl reactions containing 1µl filtered supernatant with Taq polymerase according to manufacturer's instructions, using a thermocycler protocol of 25-30 cycles of denaturation at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 1 min.

3.2.7 Plaque assays

Plaque generation was attempted using a variety of phage samples, indicator strains, and growth media. To test for spontaneous phage release, filtered supernatants of saturated cultures

were used. To test for mitomycin C-induced phage release, filtered supernatants of PEG-precipitated samples from mitomycin C-treated cultures (as generated for flow cytometry) were used. Confluent lawns were prepared by mixing 4.5 ml Reinforced Clostridial Top Agar (30 ml Reinforced Clostridial Agar + 60 ml RCM, with or without 2 mM CaCl₂) with 200-300 µl saturated bifidobacterial culture (grown overnight directly from a freezer stock) and allowed to solidify on RCA plates. For each phage sample, 3-5 µl was spotted onto the overlay and allowed to dry, and plates were incubated at 37°C in an anaerobic chamber for 24-48 h. Phage samples were generated from several lysogens (*B. choerinum* LMG 10510 and *B. moukalabense* DSM 27321) and predicted lysogens (*B. breve* 082W4-8, *B. breve* 180W8-3, *B. breve* 139W4-23, *B. breve* 017W4-39, and *B. breve* 215W4-47a), and they were tested against all of the originating lysogens and predicted lysogens as well as several non-lysogens (*B. breve* JCM 7017, *B. breve* NCIMB 702258, and *B. breve* UCC2003). As an alternative to spotting, some saturated cultures were directly mixed with phage samples in a 1.5 ml tube and aerobically incubated on the bench at room temperature for 10-15 min prior to being mixed with top agar and poured as an overlay. Additionally, TOS medium (Sigma) was used as an alternative to RCM.

3.2.8 Induced phage genome sequencing

DNA from 2 ml filtered culture supernatant (described above) was extracted for sequencing by incubating with 4 µl DNase I at room temperature for 1 h, then proceeding with the Norgen Phage DNA Extraction Kit according to the manufacturer's protocol. Gabriele Lugli and Marco Ventura sequenced the DNA sequenced using Illumina MiSeq technology (GenProbio, Parma, Italy) and used the MEGAnnotator pipeline for *de novo* assembly (Lugli et al., 2016a). Other than the phage genomes, the only other assembled DNA molecule observed

was a 6.5 kb plasmid in the *B. breve* 082W4-8 sample. Note: the sequencing protocol has been provided by Gabriele Lugli and Marco Ventura. For Rin shufflon variant analysis, sequencing data were analyzed with Newbler assembler and the Consed 454ContigGraph output.

3.2.9 Induced phage replication quantification

Sequencing reads were trimmed at both ends with Cutadapt (<https://cutadapt.readthedocs.org>) using the quality score option and a value of 30. Trimmed reads were mapped with Bowtie2 (Langmead and Salzberg, 2012), all non-unique reads were discarded using *sed*, and the data was processed with SAMtools (Li et al., 2009) and BEDtools (Quinlan and Hall, 2010). Reads were mapped to the published lysogen sequence and visualized with Integrative Genomics Viewer (Thorvaldsdottir et al., 2013). To quantify enrichment of the induced phage relative to the host, average coverage per genome was computed by dividing the number of base pairs mapped by the total size of the host or prophage genome, and fold increase in coverage was computed by dividing the average coverage of the prophage genome by the average coverage of the host genome.

3.2.10 Transmission electron microscopy

Negative straining of phage particles in filtered, mitomycin C-treated, culture supernatants (described above) was performed by Charles M.A.P. Franz and Horst Neve using freshly prepared ultra-thin carbon films with 2% (w/v) uranyl acetate as previously described (Casey et al., 2014). Micrographs were taken using a Tecnai 10 transmission electron microscope (FEI Thermo Fisher, Eindhoven, The Netherlands) at an acceleration voltage of 80 kV with a

MegaView G2 CCD-camera (emsis, Muenster, Germany). Note: this protocol has been provided by Charles M.A.P. Franz and Horst Neve.

3.2.11 Flow cytometry sample preparation and processing

Sample preparation and processing for flow cytometric analysis was performed similarly to previously described methods (Oliveira et al., 2017). Strains were grown in RCM to early log phase, treated with mitomycin C (or were left untreated), and filtered (described above). Paired treated and untreated samples were processed in parallel. 25 ml of filtered supernatant of treated/untreated cultures were incubated with 2.5 g PEG8000 on a shaker overnight at 4°C and spun in a Sorval centrifuge at 17,620 x g for 15 min at 4°C. The supernatant was discarded, and the pellets were resuspended with 1 ml TBT buffer and transferred to a 1.5 ml tube. Samples were processed by spinning in a microcentrifuge at 10,000 x g for 4 min, washed twice with 1 ml ¼ strength Ringer's solution, incubated at room temperature for 30-60 min, washed once more, and resuspended in 1 ml ¼ strength Ringer's solution. Pellets ranged in size and opacity across strains, and they were diluted 1:10 or 1:100 with ¼ strength Ringer's solution as needed for FACSCalibur flow cytometry. Using the Live/Dead BacLight Kit (Thermo Fisher), 100 µl of sample was diluted with 888.5 µl ¼ strength Ringer's solution, incubated at room temperature in the dark for 15 min with 1.5 µl Syto9 dye, and spiked with 10 µl microsphere bead standards. Samples were processed with a FACSCalibur. Forward scatter (FSC-H), side scatter (SSC-H), and fluorescence (FL1-H) parameters were measured using instrument settings that were calibrated with mitomycin C-treated *Lactococcus lactis* UC509.9 and NZ9000(TP901-1) samples prepared by Joana Oliveira to reproduce her results reported previously using a different flow cytometer (Oliveira et al., 2017). Several types of controls were used for downstream analysis,

including distilled H₂O, ¼ strength Ringer's solution, and ¼ strength Ringer's solution with beads, with and without mitomycin C added. For each sample, 100,000 events were analyzed at a rate of ~ 3,500-5,000 events/second. All strains were grown in RCM for direct comparison, although this medium produces higher flow cytometry background than other growth media.

3.2.12 Flow cytometry data analysis

FACSCalibur data were analyzed with R (version 3.4.2) (<http://www.R-project.org>) using RStudio (version 1.0.153) (<http://www.rstudio.com>) and the *flowCore* (Hahne et al., 2009) and *flowWorkspace* (Greg Finak, 2011) packages. Flow cytometric analyses of different phage types have shown that since the standard 488 nm wavelength is larger than the average phage particle, forward and side scatter parameters do not correlate with phage size (Brussaard et al., 2000). Also, fluorescence intensity of stained particles does not correlate with genome size (Brussaard et al., 2000). Therefore, flow cytometry events due to debris or bead standards were gated and removed similarly to previously described methods (Oliveira et al., 2017). To define the gates, the signal distribution of each parameter (FSC-H, SSC -H, FL1-H) was analyzed in several control samples to identify the signal range associated with each event type. This resulted in debris event boundaries of FSC-H (-Inf, 50), SSC-H (-Inf, 100), and FL1-H (-Inf, 15) and bead event boundaries of FSC-H (150, 1000), SSC-H (800, 2700), and FL1-H (15, 90). Three dimensional gates using these boundaries account for nearly all debris or bead events in the control samples. For all test samples, events passing through either the debris or the bead gates are removed, and the remaining “gated” events are used for downstream analysis of phage induction. For each paired (treated/untreated) sample, the fluorescence intensity of gated events and the ratio of gated events to total events were quantified. To assess patterns of induction,

changes in the gated/total event ratio and the median fluorescence were computed using replicate data either for each strain or for each strain type. Statistical significance was computed with the two-tailed *t*-test function in R.

3.2.13 Rin shufflon analysis

Tyrosine recombinases were analyzed with HHpred using the *PDB_mmCIF70* database, and representative domain hits were chosen to illustrate the approximate domain boundaries (N-terminal arm-type DNA-binding = 3JU0_A, 3JTZ_A; common core DNA-binding = 2OXO_A, 3NRW_A, 3LYS_B; catalytic = 5DOR_A, 1AE9_A, 5DCF_A). A phylogenetic tree of recombinases was constructed using maximum likelihood from a codon alignment by webPRANK (Loytynoja and Goldman, 2010). A stop codon is present in the middle of the Bl30698phi1 *rin* gene due to a point mutation or sequencing error. For phylogenetic purposes, the point mutation was changed to match the other alleles and the full-length gene was analyzed. For *Rv* analysis, the nucleotide sequences of all potential full-length *Rc-Rv* alleles were created, in which *Rc* was fused to each separate *Rv* sequence using the identified upstream *rix* site as the point of fusion. Full length translations were analyzed with HHpred using the *PDB_mmCIF70* database. The N-terminus *Rc* region exhibits similarity to the RBP of *L. lactis* phage 1358 (domain hit 4L9B_A). The C-termini of all *Rc-Rv* protein fusions, except for *Rc-Rv5*, exhibit similarity to the RBP of *L. lactis* phage TP901-1 (domain hits 4IOS_A, 4HEM_C, 2F0C_A). Pairwise amino acid sequence similarities of the variable C-terminus were computed using the EMBOSS Needle global alignment tool (Rice et al., 2000).

3.2.14 Rin shufflon analysis in WGS reads

Genomic inversions within the Rin shufflon of the uninduced Bb423phi1 prophage were identified by Francesca Bottacini in previously reported *B. breve* 139W4-23 raw whole genome sequencing reads (Bottacini et al., 2017). PacBio long reads (average read length > 10 kb) that map across the Rin shufflon locus (coordinates 1,307,900-1,310,888) with at least 80% sequence identity were selected using GLAT aligner v36.2. Variant shufflon orientations in this subset of reads were identified using dot plot alignments in MUMmer v3.0. Sequence coverage of each variant was computed using the identified long reads as reference sequences and performing a mapping assembly using the RS_Resequencing.1 protocol implemented in SMRT Analysis portal v2.3. The resulting assembled reads were inspected using Tablet (<https://ics.hutton.ac.uk/>). Note: this protocol has been provided by Francesca Bottacini.

3.2.15 *dnaJ₂*-integrating phage attachment site analysis

For the induced *dnaJ₂*-integrated prophages, the 7 bp point of strand exchange was determined by aligning the *attL* and *attR* sites with the *attP* site from the induced virion genome. Using the point of strand exchange, the theoretical *attB* and *attP* sites in *B. breve* and *B. longum* host strains and virion genomes were created for all other *dnaJ₂*-integrated prophages.

3.2.16 Host 16S rRNA analysis

The annotated 16S rRNA genes from bifidobacterial genomes were aligned with MUSCLE. The alignment was trimmed at both ends using CLC Genomics (CLC bio-Qiagen, Aarhus, Denmark), and a phylogenetic tree was constructed using the BioNJ algorithm in SeaView (Gouy et al., 2010).

3.2.17 Gene content flux analysis

Changes in gene content and nucleotide sequence similarity were computed as described in Chapter 2. Evolutionary modes were predicted as described in Chapter 2.

3.3 RESULTS

3.3.1 Bioinformatic characterization of bifidophages

A previous investigation of several *Bifidobacterium breve* isolates (082W4-8, 180W8-3, 139W4-23, 017W4-39, 215W4-47a) bioinformatically identified several potential prophages (Bottacini et al., 2017), designated here as Bb48phi1, Bb83phi1, Bb423phi1, Bb423phi2, Bb439phi1, and Bb447phi1. BLAST searches using their annotated integrase genes identified potential prophages in five other isolates, including *B. breve* JCM 1192, *B. breve* 689b (Bottacini et al., 2014), *B. longum* subsp. *infantis* ATCC 15697 (Fukuda et al., 2011), *B. longum* subsp. *longum* 157F (Fukuda et al., 2011), and *B. longum* subsp. *longum* CCUG 30698 (O'Callaghan et

al., 2015). Two of these prophages, 689b-1 (Bottacini et al., 2014) and Binf-1 (Ventura et al., 2009) have been reported previously, while the other three are newly identified and designated here as Bb1192phi1, B1157phi1, and B130698phi1. In order to bioinformatically assess the potential for these prophages to form infectious particles, their genomic architectures were characterized and compared to isolated actinobacteriophages using Phamerator (Cresawn et al., 2011), which identifies regions of nucleotide sequence similarity and groups gene products with similar sequence into phamilies (“phams”).

Seven prophages are integrated at the *dnaJ₂* locus (Figure 3-1A). DnaJ₂ is one of two DnaJ homologs present in the *Bifidobacterium* genome. DnaJ₂ is a highly conserved molecular chaperone involved in stress response, similar to DnaJ₁, but it is only present in the Actinobacteria phylum (Ventura et al., 2005a). For each prophage, the 35 bp *attL* overlaps the 3’ end of the *dnaJ₂* gene, and the *attR* is flanked by the integrase gene. Integration into *dnaJ₂* forms a complete coding sequence such that the encoded amino acid sequence is unaffected. The prophages range in size from 39-43 kb, contain genes associated with several stages of phage growth (including integration, transcriptional regulation, replication, particle assembly, and host lysis), and two (Bb83phi1 and Bb439phi1) are nearly identical. The *dnaJ₂*-integrated phages share five phams with their closest relative in the actinobacteriophage database, *Streptomyces* phage phiSASD1, which was originally isolated from a fermentation factory and may be temperate (Figure 3-2A)(Wang et al., 2010a). The shared phams are syntenically positioned and are associated with particle structure and assembly. These are the only structural and assembly genes shared between the *dnaJ₂*-integrated prophages and the 54 *Streptomyces* phages in the actinobacteriophage database, indicating that the similarities to phiSASD1 are not a general pattern with other types of *Streptomyces* phages.

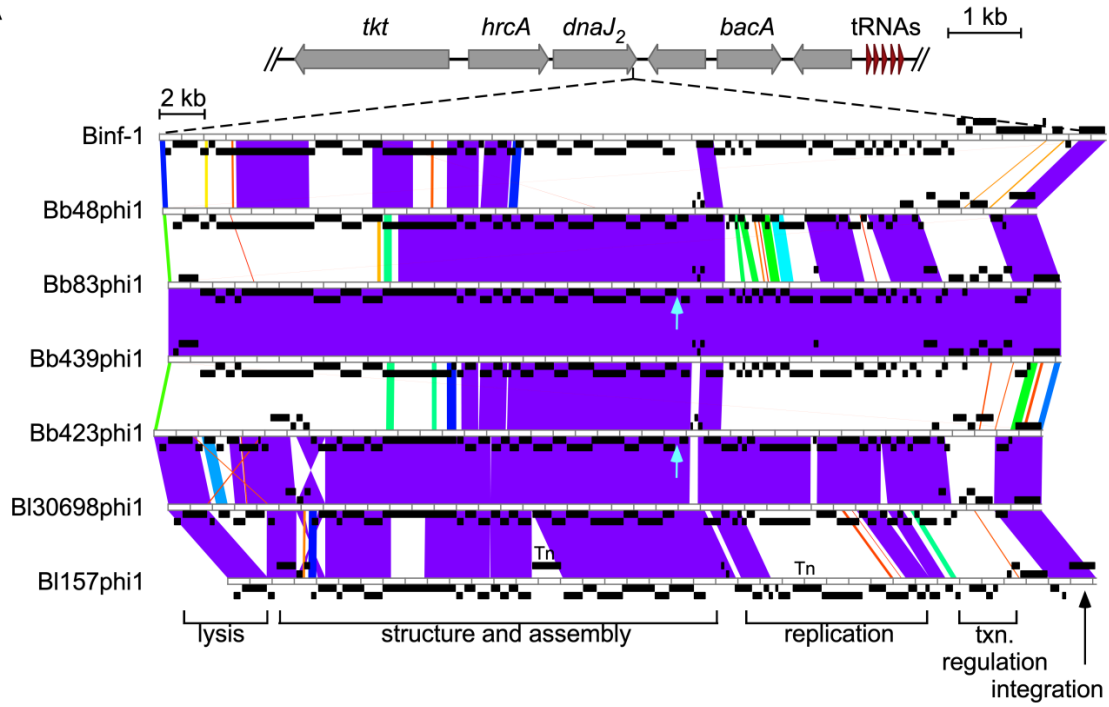
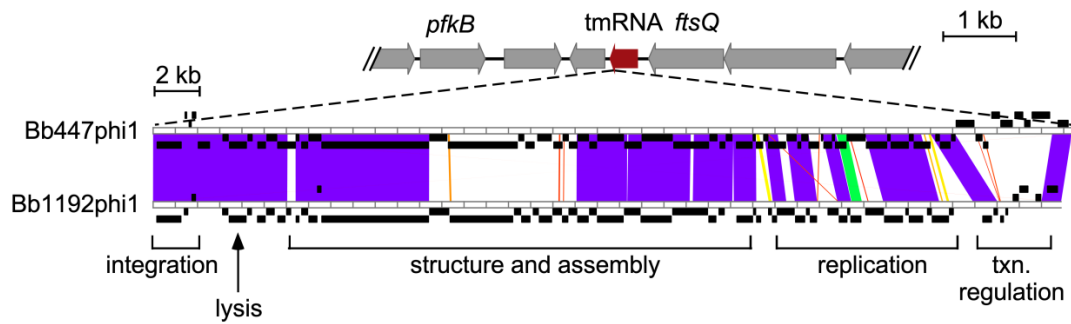
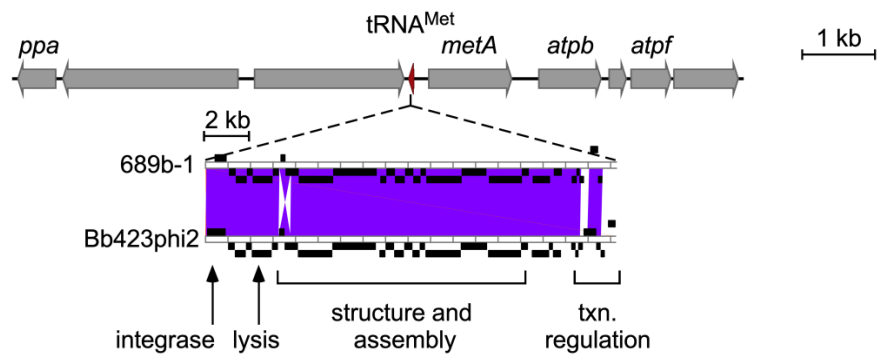
A**B****C**

Figure 3-1. Bifidoprohage genomic comparison and characterization.

(A) (top) Enlarged view of the *dnaJ₂* integration locus. Coding (grey) and tRNA (red) genes are indicated, oriented in the direction of transcription, and with gene descriptions indicated where applicable. The seven prophages are integrated at the 3' end of the *dnaJ₂* gene. (bottom) Genome architecture and mosaic relationships between the seven prophage genomes are highlighted with pairwise alignments in Phamerator. Genes (black boxes) are positioned above or below the genome ruler to indicate orientation. The color spectrum between genomes reflects sequence similarity based on BLAST *e*-values, ranging from white (no similarity) to violet (high similarity). Cyan arrows indicate the areas of lowest sequencing coverage from the induced virion genomes and the likely locations of the linear virion genome termini. General regions of specific gene modules are indicated below the alignment. Tn = transposase. **(B)** (top) Enlarged view of the tmRNA integration locus, as in panel A. The prophages are integrated at the 3' end of the tmRNA gene. (bottom) Genome architecture and mosaic relationships between the prophage genomes, as in panel A. **(C)** (top) Enlarged view of the tRNA^{Met} integration locus, as in panel A. The prophages are integrated at the 3' end of the tRNA^{Met} gene. (bottom) Genome architecture and mosaic relationships between the prophage genomes, as in panel A. Note: in each panel, the host and prophage genome maps are on different scales. Figure adapted from (Mavrich et al., 2018).

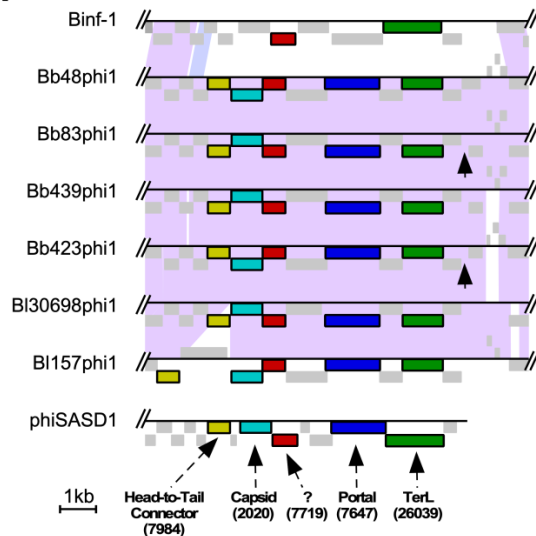
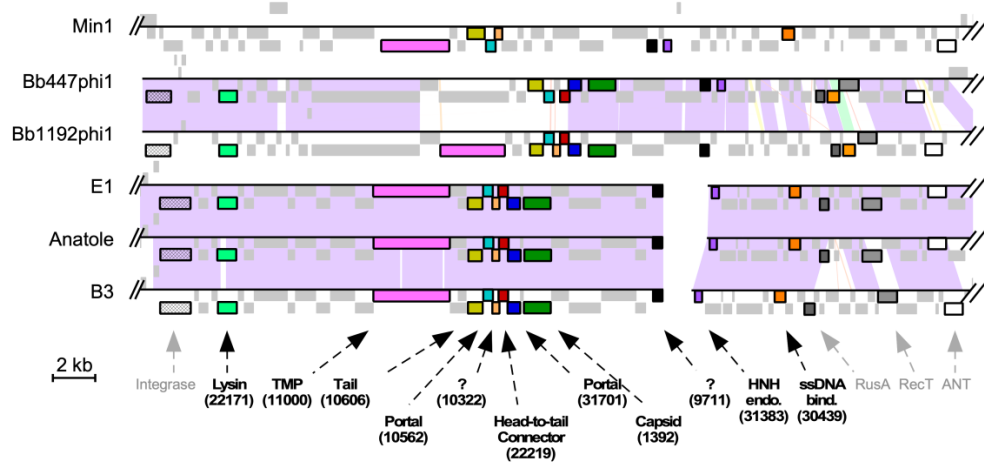
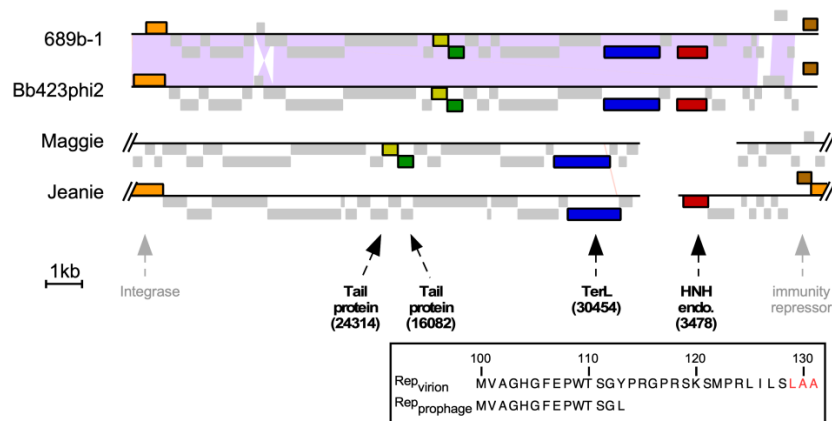
A**B****C**

Figure 3-2. Evolutionary relationships of bifidophages to other actinobacteriophages.

(A) Enlarged view of the structural gene locus of *dnaJ₂*-integrated prophages, from Figure 3-1A. *Streptomyces* phage phiSASD1 (in virion orientation) has been included for comparison. Genes of all phams that are shared between phiSASD1 and at least one of the *dnaJ₂*-integrated prophages are highlighted. Each pham is uniquely color-coded and labeled with the predicted function and pham number. All other genes are grey. Black arrows indicate the location of the linear virion genome termini. **(B)** Enlarged view of tmRNA-integrated prophages from Figure 3-1B. *Microbacterium* phage Min1 and *Propionibacterium* phages E1, Anatole, and B3 (in virion orientation) have been included for comparison. Genes of all phams that are shared between a *Propionibacterium* or *Microbacterium* phage and at least one tmRNA-integrated prophage are highlighted. Each pham is uniquely color-coded and labeled in bold with the predicted function and pham number. Genes that are present in all genomes and that have a similar predicted function, but are not in the same pham, are also highlighted. Each gene is uniquely color-coded and labeled with the function in grey, but pham numbers are omitted. All other genes are grey. **(C)** Enlarged view of tRNA^{Met}-integrated prophages from Figure 3-1C. *Arthrobacter* phage Maggie and *Gordonia* phage Jeanie (in virion orientation) have been included for comparison. Genes colored and indicated as in panel B. (Inset) C-terminal amino acid sequence of the Bb423phi2 immunity repressor (Rep) translated from the sequenced prophage and predicted virion isoforms; with an *ssrA*-like ClpX recognition motif highlighted (red). Figure adapted from (Mavrich et al., 2018).

Prophages Bb447phi1 and Bb1192phi1 are integrated at a tmRNA locus (Figure 3-1B). Each are 41 kb in size with a 26 bp *attR* that overlaps the 3' end of the tmRNA gene and an *attL* that is flanked by the integrase gene. They contain genes associated with several stages of phage growth similar to the *dnaJ₂*-integrated prophages. The closest relatives in the actinobacteriophage database are *Microbacterium* phage Min1 (Akimkina et al., 2007) and *Propionibacterium* phages E1, Anatole, and B3 (Cheng et al., 2018)(Figure 3-2B). These phages share 10-11 syntenically positioned phams, many with predicted functions related to genome replication, particle structure and assembly, and host lysis. Additionally, they contain several other syntenically positioned genes with similar predicted functions (such as RusA, RecT and ANT), even though they are more distantly related and grouped into different phams.

Prophages 689b-1 and Bb423phi2 are integrated at the tRNA^{Met} locus (Figure 3-1C). They contain a 39 bp *attR* that overlaps the 3' end of the tRNA^{Met} gene and is flanked by the

integrase gene. In contrast to prophages at the other two loci though, these genomes are only 20 kb in size and lack a recognizable operon associated with genome replication. Due to the small size, 689b-1 was previously predicted to be a cryptic prophage (Bottacini et al., 2014). However, these prophages exhibit gene content similarity with isolated Cluster AN phages infecting *Arthrobacter* hosts (Klyczek et al., 2017), such as Maggie, and Cluster CW phages infecting *Gordonia* hosts (Pope et al., 2017), such as Jeanie, all of which have small genomes (Figure 3-2C). Similar to Jeanie and its relatives, the bifidoprophages have a transcriptional regulator gene flanking the *attR* (Figure 3-2C), similar to characterized integration-dependent immunity systems (Broussard et al., 2013; Pope et al., 2017). Phages with this system express two isoforms of the immunity repressor. During lytic growth, a longer isoform of the repressor is expressed, but it is unstable since the C-terminal end contains an *ssrA*-like tag that targets the protein for degradation. The *attP* site is located within the repressor gene, and after integration a stable, shorter isoform lacking the degradation tag is expressed since the repressor gene becomes truncated. Reconstruction of the two bifidoprophage virion sequences from the identified attachment sites reveals that they may also encode a second, longer repressor isoform during lytic growth (Figure 3-2C). The virion isoform of each putative immunity repressor contains an additional C-terminal 18 amino acids that contain an *ssrA*-like ClpX recognition motif. Thus, despite their small genome size, these bifidoprophages may be capable of forming phage particles.

Genetically related prophages in different bifidobacterial isolates may be the result of multiple, independent phage infections, or they may represent derivatives of a single ancestral integration event. In general, the prophages exhibit similar GC% content as their hosts (Tables 3-1, 3-2), and prophages integrated at the same site exhibit higher nucleotide sequence similarity to

each other than to prophages at other sites (Figure 3-3A). However, they do not correlate with genetic differences in the host 16S rRNA (Figure 3-3B). For instance, although *B. breve* JCM 1192 is closely related to *B. breve* 139W4-23, it lacks prophages at the *dnaJ₂* and tRNA^{Met} loci but contains a prophage at the tmRNA locus. Additionally, the prophages exhibit sequence similarity to previously reported bifidoprophages that are derived from diverse bifidobacterial host species (Figure 3-3C, analysis performed by Gabriel Andrea Lugli and Marco Ventura). Lastly, gene content in prophages may be impacted by recombinational events, such as for Bl157phi1, which has acquired two predicted transposases, *BLIF_1064* and *BLIF_1048* (Figure 3-1A). Each transposase is 1,399 bp in size, contains a single gene (the transposase), and is flanked by 50 bp inverted repeats. It is not clear whether the two transposases interfere with phage replication and growth after induction, but they are nearly identical to each other and to transposases in the host genome, so they may be recent recombinational events. However, none of the other six prophages at the *dnaJ₂* locus contain these transposases. Overall, although similar types of prophages correlate with distinct integration sites, at least some of the genetic diversity is likely the result of separate infection and integration events.

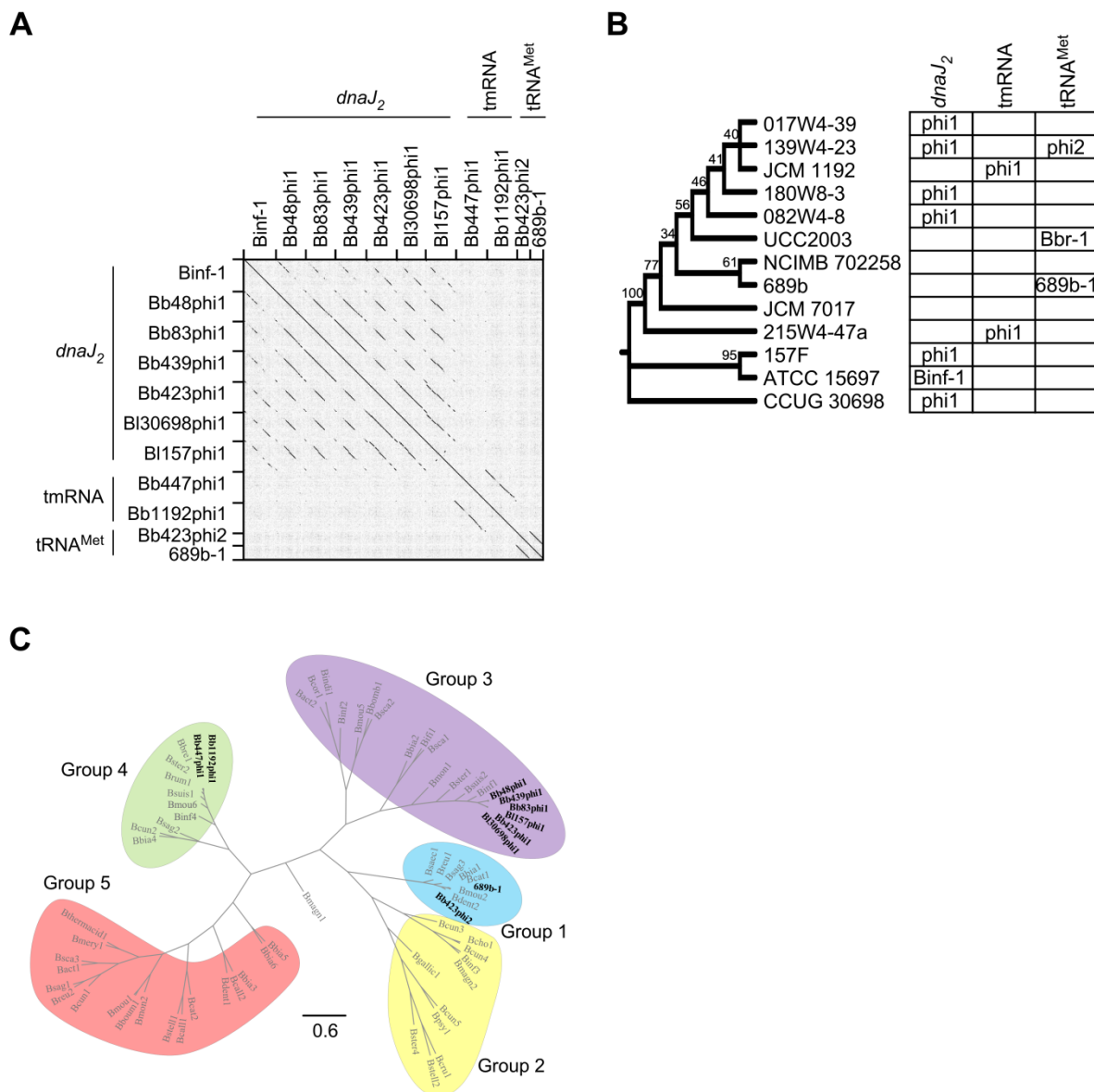


Figure 3-3. Genomic relationships of bifidoprophages and their hosts.

(A) Gepard dot plot analysis highlights prophage pairwise sequence similarities. (B) Cladogram of bifidobacterial host genomes constructed from alignment of 16S rRNA sequences; % bootstrap branch support indicated. Table indicates the presence or absence of prophages integrated at the *dnaJ₂*, tmRNA, or tRNA^{Met} loci. (C) Phylogenetic analysis of sixty previously reported bifidoprophages (grey)(Lugli et al., 2016b) and eleven newly characterized bifidoprophages (black) using whole genome alignments and grouped as previously described (Lugli et al., 2016b). Note: this analysis was performed by Gabriel Andrea Lugli. Figure adapted from (Mavrich et al., 2018).

3.3.2 Prophages can be induced with mitomycin C

The genomic characteristics of these prophages suggest that some of them may be able to produce infectious phage particles. Lysogens may spontaneously release infectious phage particles during growth, which can be determined with plaque assays using sensitive host strains. However, plaque assays using a variety of lysogens, indicator strains, media, and growth conditions failed to produce infectious particles (see Materials and Methods), similar to reports with other bifidobacterial lysogens (Lugli et al., 2016b; Ventura et al., 2005b). The absence of plaques may indicate that the prophages are not capable of producing infectious particles or that no permissive hosts are available.

Mitomycin C treatment is an effective strategy to induce prophages, including those from lysogens *B. choerinum* LMG 10510 and *B. moukalabense* DSM 27321 (Lugli et al., 2016b). Therefore, the impact of mitomycin C on the growth of five *B. breve* strains was investigated. First, the appropriate concentration of mitomycin C needed to induce prophages was determined by measuring growth of non-lysogens and predicted lysogens in the presence of a range of mitomycin C concentrations using 96-well microtiter plates (see Materials and Methods). Growth inhibition was observed for non-lysogens above 0.3 µg/ml (data not shown), so this concentration was used to test all strains, similar to previously reported methods (Lugli et al., 2016b).

The level of growth inhibition due to mitomycin C treatment was determined by comparing culture densities of treated and untreated samples for three non-lysogens [*B. breve* JCM 7017 (Bottacini et al., 2014), *B. breve* NCIMB 702258 (Bottacini et al., 2014), and *B. breve* UCC2003 (O'Connell Motherway et al., 2011)], two lysogens [*B. choerinum* LMG 10510 and *B. moukalabense* DSM 27321 (Lugli et al., 2016b)], and five predicted *B. breve* lysogens (Bottacini

et al., 2017) (Figure 3-4). In the absence of mitomycin C, strains grow to similar saturation densities, with the exception of *B. breve* 215W4-47a. After mitomycin C treatment, the lysogens and some predicted lysogens (*B. breve* 180W8-3, *B. breve* 139W4-23, and *B. breve* 017W4-39) generally exhibit more substantial growth inhibition than non-lysogens. However, since inhibition of non-lysogenic strains is variable, growth inhibition of the predicted lysogens might not be the result of prophage induction.

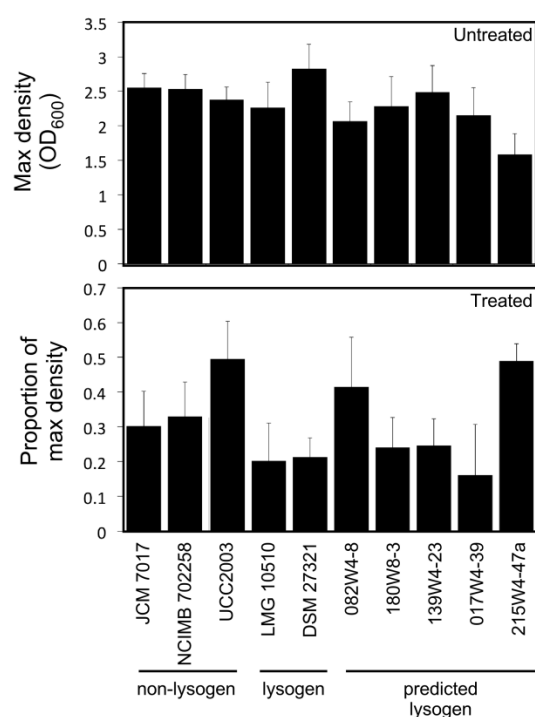


Figure 3-4. Impact of mitomycin C on bifidobacterial growth.

Growth characteristics for several non-lysogens, lysogens, and predicted lysogens that were (top) untreated or (bottom) treated with mitomycin C. (top) Bar plot displays the average maximum saturated culture density (based on OD_{600nm}). (bottom) Bar plot displays the effect of mitomycin C treatment on maximum culture density. Cultures were treated with 0.3 µg/ml mitomycin C at OD_{600nm} ~ 0.15-0.25 and after overnight incubation the final density was measured. An untreated sample was grown overnight as well, and the ratio of the treated versus untreated maximum saturated culture density was determined. Error bars indicate standard deviation from three or more replicates. Figure adapted from (Mavrich et al., 2018).

3.3.3 Phage genomes circularize after induction

Successful induction is expected to result in the excision and circularization of the prophage genome. The presence of excised virion genomes in mitomycin C-treated samples was measured by PCR amplification across the predicted *attP* locus in unfiltered or filtered supernatants from saturated cultures (Figure 3-5A). Excision and circularization are detected for *dnaJ*₂-integrated Bb439phi1 and Bb83phi1 as well as tmRNA-integrated Bb447phi1 after mitomycin C, similar to the expected excision of the prophage in *B. choerinum* LMG 10510 (Figure 3-5B). Circularization of *dnaJ*₂-integrated Bb48phi1 and Bb423phi1 phages is detected regardless of mitomycin C treatment, suggesting there may be low levels of spontaneous prophage excision. In contrast, circularization for Bb423phi2 at the tRNA^{Met} locus was not detected. Thus, many of these prophages are capable of excising and circularizing, either spontaneously or as a result of mitomycin C treatment.

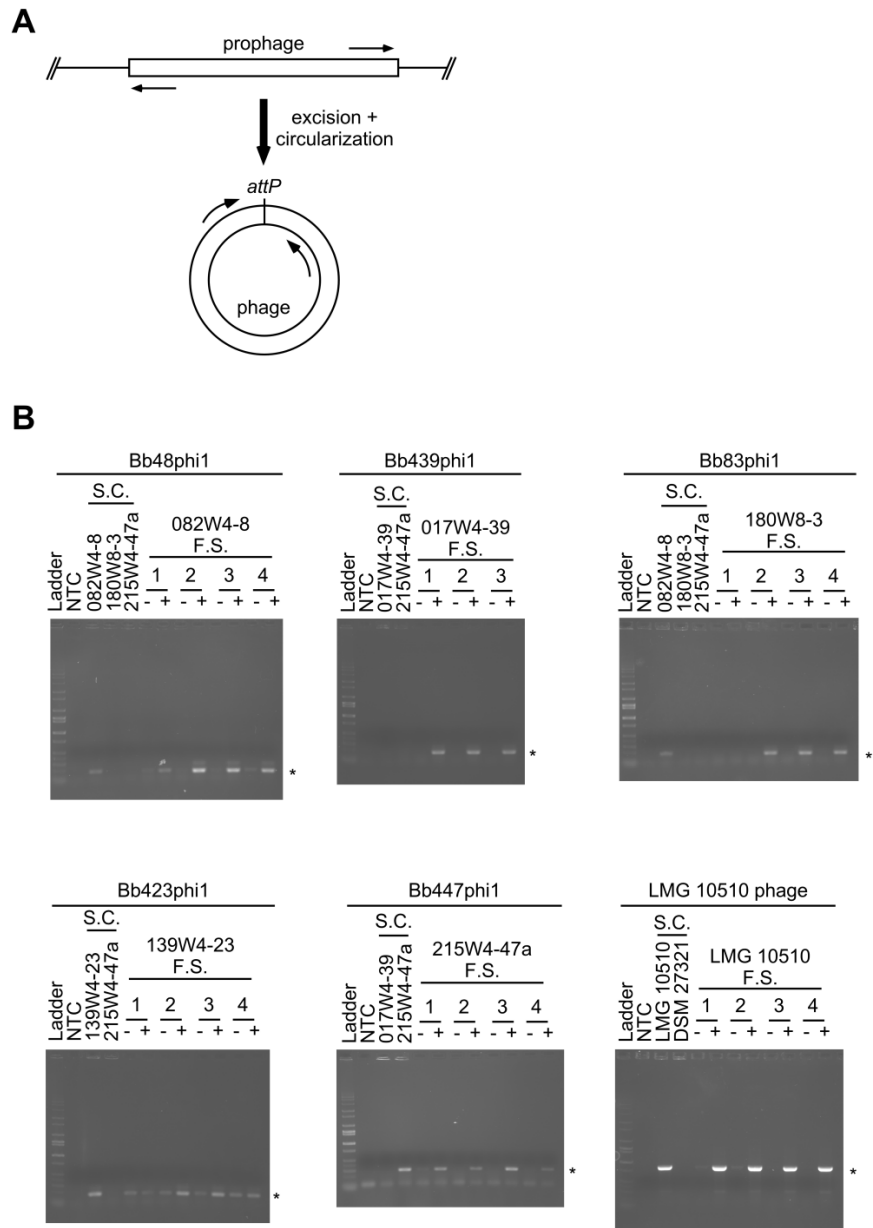


Figure 3-5. Mitomycin C induces prophage excision and circularization.

Excision and circularization of the predicted prophages were examined by PCR. **(A)** Primers (arrows) were designed in all predicted prophages so that they are divergent in the integrated genome orientation but convergent in the excised and circularized genome to amplify across the *attP*. **(B)** Prophage induction was tested in several strains by PCR amplification of filtered culture supernatants (F.S.) treated (+) or not treated (-) with mitomycin C. Three to four replicates were tested per strain. For each prophage of interest, a no template control (NTC) and several unfiltered saturated cultures (S.C.) were included as negative and positive controls. The full length of each lane from the loading well to leading edge is displayed. A star (*) indicates the expected band size corresponding to *attP* amplification. Figure adapted from (Mavrich et al., 2018).

3.3.4 Induced phages increase in copy number

Following induction and excision, the phage genome is expected to replicate, increasing copy number relative to the host genome. To measure replication, DNA from mitomycin C-treated culture supernatants was extracted and sequenced for cultures of *B. breve* 139W4-23, 082W4-8, 180W8-3, 017W4-39, and 215W4-47a (sequencing performed by Gabriel Andrea Lugli and Marco Ventura). Sequencing reads were mapped to the lysogen genome (Figure 3-6). In contrast to the tmRNA-integrated Bb447phi1 or tRNA^{Met}-integrated Bb423phi2 genomes, the *dnaJ*₂-integrated prophage genomes exhibit a substantial increase in sequencing coverage relative to the host genome (Figures 3-6, 3-7). Additionally, the virion genomes induced from the *dnaJ*₂ locus were assembled with 20x-200x coverage (Figure 3-7). Thus, following mitomycin C treatment, *dnaJ*₂-integrated prophages are likely replicating after excision.

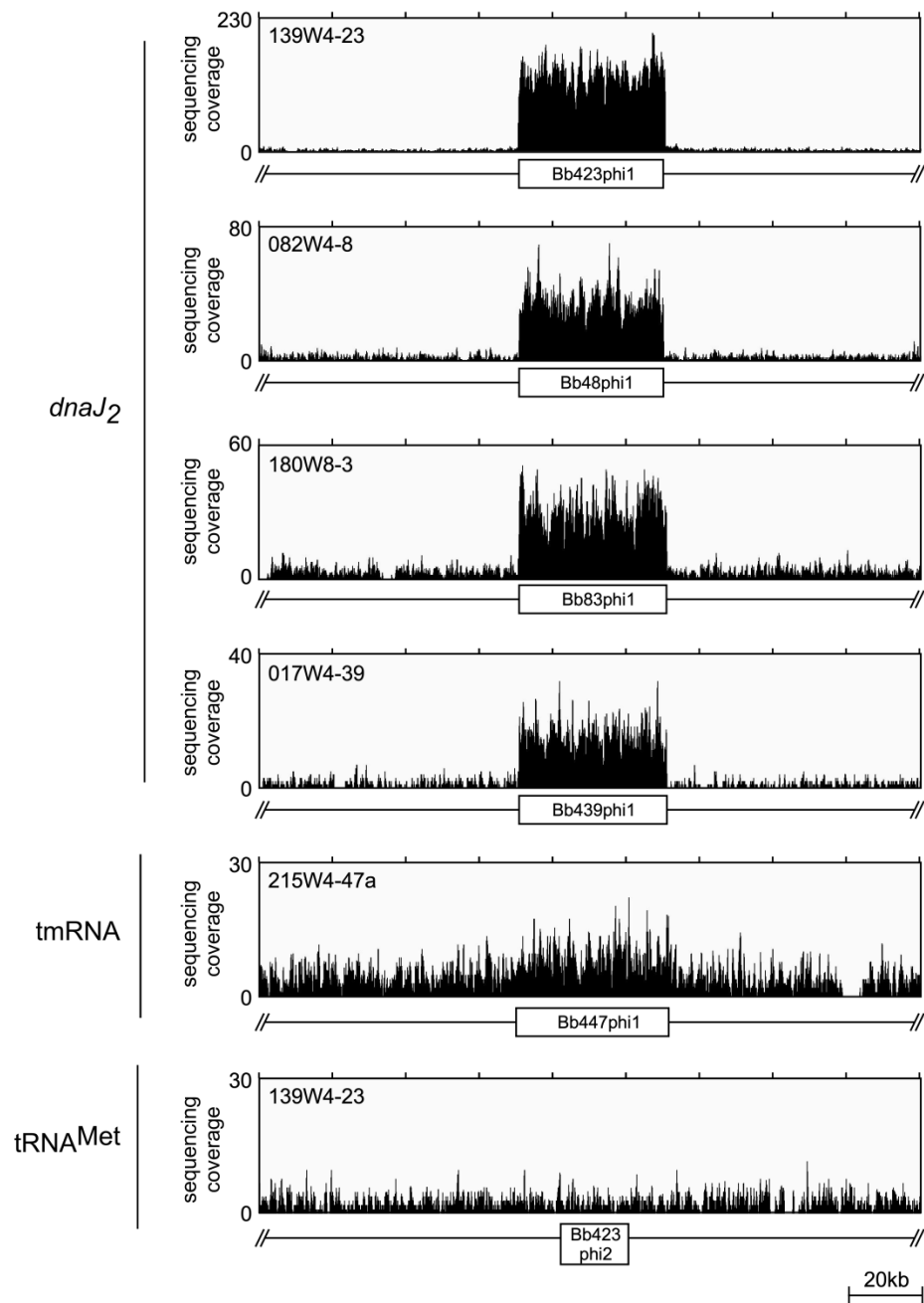


Figure 3-6. Mitomycin C increases sequencing coverage of *dnaJ2*-integrated prophages.

DNA from mitomycin C-treated culture supernatants were sequenced for several *B. breve* strains and reads were mapped to the lysogen genome (black line). Enlarged view of the integrated prophage (white box) locus in each strain highlights the increased sequencing coverage of the prophage relative to the host genome. Figure adapted from (Mavrich et al., 2018).

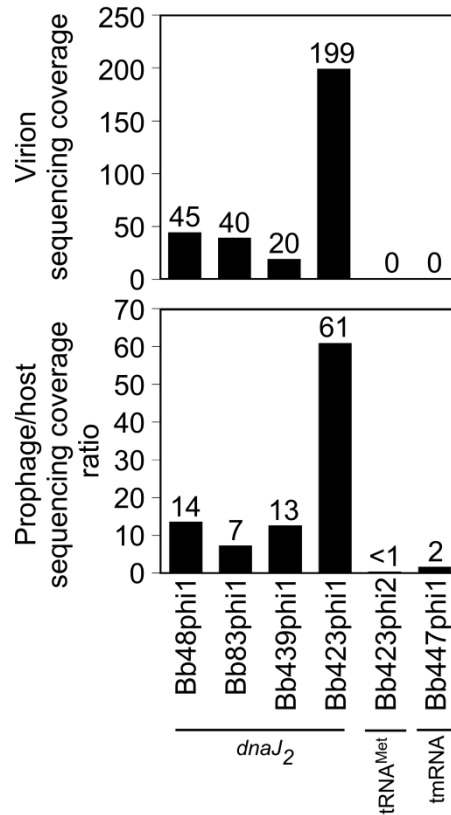


Figure 3-7. Mitomycin C increases *dnaJ₂*-integrated prophage copy number.

For several predicted lysogens, DNA was extracted from mitomycin C-treated culture supernatants and sequenced. Sequencing reads were used to (bottom) determine enrichment of the prophage genome relative to the host genome after mapping reads to the lysogen genome in Figure 3-6, and (top) assemble virion genomes and compute coverage depth. Figure adapted from (Mavrich et al., 2018).

During packaging, the circularized phage genome is linearized by Terminase. Sequencing of the linearized genome can identify the genomic position where the DNA is processed by the Terminase. For Bb83phi1 and Bb423phi1, a prominent drop in coverage is observed within the same intergenic loci upstream of the structural and assembly genes (Figure 3-1A) and in the same approximate position of the genome terminus for phage phiSASD1 (Figure 3-2A). These results suggest that at least in these two bifidophages the Terminase actively processes the replicated DNA for packaging.

3.3.5 Induced phages contain packaged DNA

Following replication and packaging, the phage is expected to lyse the cell, increasing the quantity of DNA-containing phage particles in the culture. Flow cytometry has been used to identify mitomycin C-induced *Lactococcus* phages with packaged DNA by characterizing and comparing absolute levels of forward scatter (FSC-H), side scatter (SSC-H), and DNA-stained fluorescence (FL1-H) of PEG-precipitated culture supernatants (Oliveira et al., 2017). Here, a similar strategy was used to quantify mitomycin C induced changes in supernatant composition to determine whether the linearized bifidophage DNA is effectively packaged.

If mitomycin C treatment results in phage replication and lysis, there is expected to be an increase in the abundance of flow cytometric events and fluorescent intensity (due to increased number of phage particles with packaged DNA) when paired treated and untreated samples are compared. Changes in supernatant composition are expected to be more substantial in lysogens and predicted lysogens than in non-lysogens. Bifidobacterial strains were grown in RCM and treated with mitomycin C (or were left untreated) during exponential growth. Culture supernatants were filtered, PEG-precipitated, stained with Syto9, spiked with microsphere bead standards, and processed through a flow cytometer. Since a different flow cytometer was used for these bifidobacterial samples than for the published *L. lactis* samples, background flow cytometric events were identified and removed using several controls. Mitomycin C-treated samples for a *L. lactis* non-lysogen (UC509.9) and lysogen (NZ9000 with integrated TP901-1 prophage) were used to calibrate the flow cytometer sample processing program (Figure 3-8A, samples prepared by Joana Oliveira). Once the program was defined, several negative controls were analyzed (Figure 3-8B, C). For each parameter of interest (FSC-H, SSC-H, and FL1-H), gates were defined to remove background debris and bead standards events (Figure 3-8D, E).

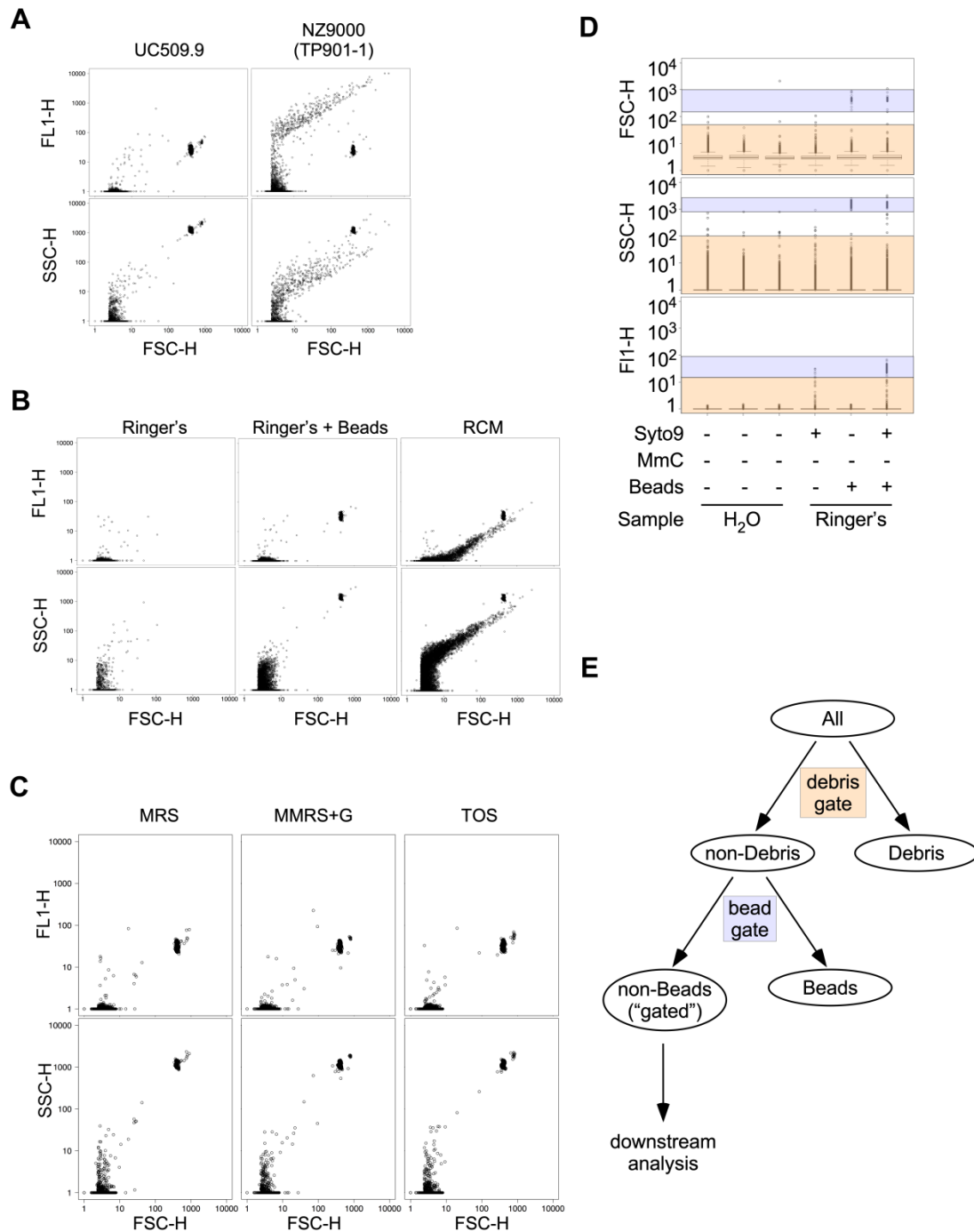


Figure 3-8. Flow cytometry calibration and gating strategy.

(A) FACSCalibur settings were calibrated using mitomycin C-treated *L. lactis* non-lysogen (strain UC509.9) and lysogen (strain NZ9000 with TP901-1 prophage) samples. Scatter plots comparing forward scatter (FSC-H) to (bottom) side scatter (SSC-H) and (top) Syto9 fluorescence (FL1-H) were adjusted to reproduce previously described results (Oliveira et al., 2017). (B) Flow cytometry of several negative controls, plotted as in panel A, to identify different types of events to gate. Samples include flow sample buffer (¼ strength Ringer's

solution), flow sample buffer with reference microsphere beads ($\frac{1}{4}$ strength Ringer's solution + beads), and growth medium processed with the entire protocol (RCM). **(C)** Flow cytometry of common bifidobacterial growth media (MRS, MMRS + Glucose, TOS), plotted as in panel B. **(D)** Boxplots of individual parameters (FSC-H, SSC-H, FL1-H) from flow cytometry results for several strain-free controls are used to define boundaries of each parameter for debris (beige) and bead (blue) events. Some samples have been treated (+) or not treated (-) with mitomycin C (MmC), Syto9 stain, and beads. **(E)** Boundaries defined in panel D were used to create three-dimensional debris and bead gates. The gating strategy for all flow cytometry samples utilizes these two gates for removal of debris events followed by removal of bead events. All non-debris and non-bead "gated" events are used for downstream analysis to assess levels of prophage induction. Figure adapted from (Mavrich et al., 2018).

Using this data analysis pipeline, all paired untreated and mitomycin C-treated bifidobacterial samples were processed. The fluorescent intensity of all non-background gated events, and the ratio of gated events to total events, were quantified for all paired samples (Figure 3-9A), and the differences in these two metrics between untreated to treated samples were assessed (Figure 3-9B). Variability is observed between replicate sets of the same strain as well as between strains of the same strain type (non-lysogen, lysogen, or predicted lysogen), suggesting that mitomycin C treatment does not reproducibly generate distinct, robust, induction-dependent changes in supernatant composition. However, despite this variability, when results of each strain type are combined, the increases in gated event abundance and fluorescent intensity for lysogens and predicted lysogens with *dnaJ₂*-integrated prophages are indeed significantly larger than increases in non-lysogenic strains (Figure 3-9C). In contrast, *B. breve* 215W4-47a, harboring a tmRNA-integrated prophage, does not exhibit the same types of changes. Thus, the *dnaJ₂*-integrated prophages exhibit significant changes in mitomycin C-dependent supernatant composition, consistent with the hypothesis that these phages package their DNA effectively.

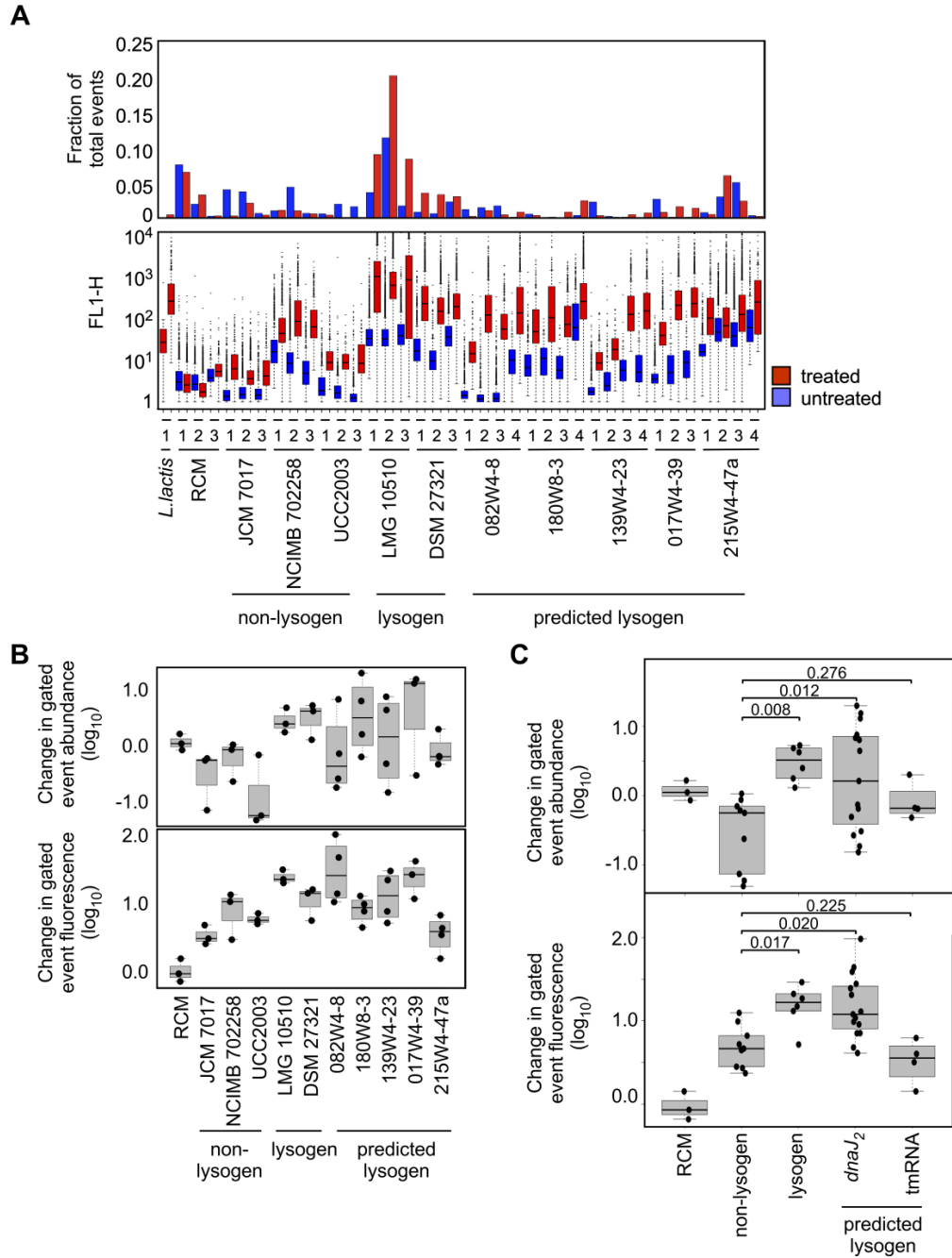


Figure 3-9. Mitomycin C induced changes in supernatant composition.

(A) For all gated events from each replicate set of paired mitomycin C treated (red) and untreated (blue) samples, (top) bar plot of the proportion of total events and (bottom) box plots of event fluorescence (FL1-H) highlight changes in supernatant composition. Replicate sets are numbered. Individual strain names are indicated along with whether they are non-lysogens, lysogens, or predicted lysogens. RCM = growth medium with no cell culture. *L. lactis* untreated = mitomycin C-treated non-lysogen (UC509.9); *L. lactis* treated = mitomycin C-treated lysogen (NZ9000 with TP901-1 prophage). (B) Box plots display fold changes in the abundance (top) and median

fluorescence (bottom) of events observed from mitomycin C treated versus untreated paired samples in panel A for growth medium (RCM) and individual non-lysogenic, lysogenic, and predicted lysogenic strains. (C) Box plots as in panel B but grouped by strain type and prophage integration loci. Black bar indicates median, and individual data points are plotted. Statistical significance of samples from different types of strains (lysogens, $n = 6$; *dnaJ*₂, $n = 15$; tmRNA, $n = 4$) compared to non-lysogens ($n = 9$) are indicated (p -value from two-tailed t -test). Figure adapted from (Mavrich et al., 2018).

3.3.6 Induction generates complete phage particles

Although the *dnaJ*₂-integrated phages exhibit mitomycin C-dependent excision, replication, packaging and lysis, it is not clear whether complete phage particles are produced. To address this, mitomycin C-treated *B. breve* 082W4-8 and *B. breve* 139W4-23 filtered culture supernatants were analyzed by electron microscopy (performed by Charles M.A.P. Franz and Horst Neve). In both samples, phage particles are observed, but they are at low concentration and sometimes contain empty capsids (Figure 3-10, Table 3-3). They are both siphoviral with approximately 200 nm tails, and particles from *B. breve* 139W4-23 contain unique tail fibers and discs present along the tail (Figure 3-10). Particles present in the *B. breve* 082W4-8 sample are likely derived from Bb48phi1 since that is the only prophage identified in this strain. Particles present in the *B. breve* 139W4-23 sample are likely derived from Bb423phi1, since Bb423phi2 induction is not detected (Figures 3-5, 3-6, 3-7).

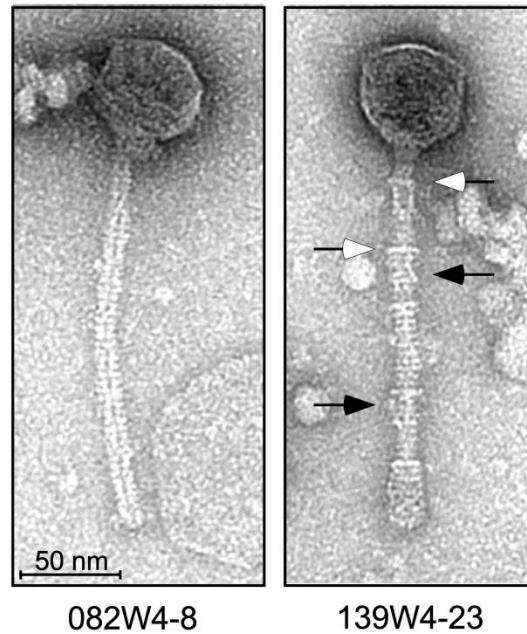


Figure 3-10. Complete phage particles are present in mitomycin C induced samples.

Transmission electron microscopy of mitomycin C-treated supernatants of *B. breve* 082W4-8 and *B. breve* 139W4-23 cultures. The phage induced from *B. breve* 139W4-23 contains tail decoration discs (open arrows) and tail decoration fibers (closed arrows). Note: experiment and figure were generated by Charles M.A.P. Franz and Horst Neve. Figure adapted from (Mavrich et al., 2018).

Table 3-3. Dimensions of bifidophages detected by TEM.

Strain	Head (nm)	Tail length (nm)	Tail width (nm)	Tail decorations width (nm)
<i>B. breve</i> 082W4-8	60.3 ± 0.04 (n=2)	200.5 ± 2.2 (n=2)	11.8 ± 0.6 (n=2)	N/A
<i>B. breve</i> 139W4-23	60.3 ± 3.3 (n=5)	193.7 ± 1.7 (n=5)	13.4 ± 0.6 (n=5)	17.1 ± 1.0 (n=8)

Note: data/table were generated by Charles Franz and Horst Neve and adapted from (Mavrich et al., 2018).

The electron microscopy results confirm that for these two strains, and possibly for other *B. breve* lysogens with *dnaJ₂*-integrated prophages, complete phages are produced. Therefore, plaque assays were performed using mitomycin C-treated supernatants of several predicted

lysogens and several indicator strains (see Materials and Methods). However, no plaques were observed again, suggesting that the inability to generate infectious particles is likely caused by inadequate host growth conditions or lack of permissive strains instead of an absence of phage particles.

3.3.7 Characterization of the Rin shufflon

Some phages contain receptor binding proteins (RBPs) on the distal end of the tail structure that specifically recognize molecules on the host's outer cell surface and confer host specificity. Although the seven *dnaJ₂*-integrated prophages contain structural and lysis genes in syntenic positions and harbor a tape measure protein (TMP) of the same pham, they do not all contain the same distal tail gene modules (Figure 3-11). Prophages Binf-1, Bb48phi1, Bb83phi1, and Bb439phi1 contain two genes immediately downstream of the TMP that exhibit distant similarity to the distal tail protein (DIT)(Bebeacua et al., 2010) and RBP (Spinelli et al., 2006) of *Lactococcus* phage TP901-1. In contrast, prophages Bb423phi1, Bl30698phi1, and Bl157phi1 contain an RBP module that is structured strikingly similar to the highly characterized phase variation loci Gin (in enterobacteria phage Mu), Cin (in enterobacteria phage P1), and Min (in the cryptic extrachromosomal enterobacteria phage p15B)(Sandmeier, 1994). These loci are composed of a single copy of a “constant” tail fiber N-terminus sequence immediately adjacent to several copies of a “variable” C-terminus sequence. An adjacent invertase inverts the orientation of the variable genes utilizing interposed crossover sites to create novel tail fiber genes. The expression of the fluctuating tail fiber sequence enables the phage to modulate host range specificity.

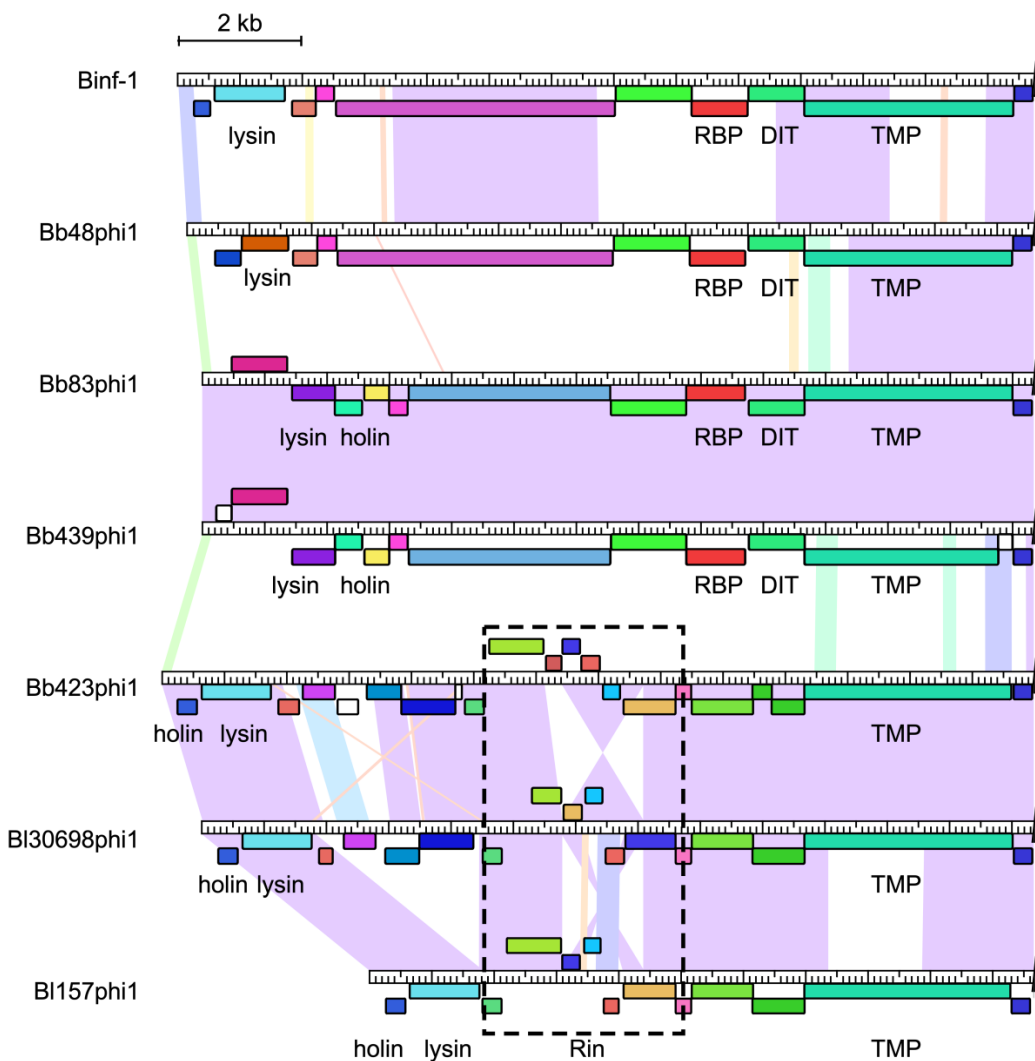


Figure 3-11. *dnaJ2*-integrated prophages contain the Rin shufflon.

Enlarged view of the left arm genes of *dnaJ2*-integrated prophages from Figure 3-1A highlights the genomic context of the Rin shufflon. Genes are colored according to pham designation, and any putative functions are listed (TMP = tape measure protein; DIT = distal tail protein; RBP = receptor binding protein). The color spectrum between genomes is the same as in Figure 3-1. Figure adapted from (Mavrich et al., 2018).

The RBP locus of the latter three bifidophages contains all of the analogous components to the Min system (Johnson, 2015; Sandmeier et al., 1992). The bifidophage locus contains several small, tandem genes, most of which exhibit distant similarity to the C-terminus of the previously characterized RBP of *Lactococcus* phage TP901-1 (Figure 3-12A). These RBP C-terminus variable (*Rv*) genes are flanked on one side by the RBP N-terminus constant (*Rc*) gene, which exhibits distant similarity to the RBP of *Lactococcus* phage 1358. On the opposite side of the *Rv* genes is a predicted recombinase, designated here as the RBP locus invertase (*rin*). Lastly, upstream of each *Rv* gene is a short, 11 bp repeated sequence (TTCCCTAACCC), likely the crossover sites (*rix*) facilitating inversion. Additionally, the *Rv* proteins are very dissimilar from each other, ranging between 6-30% amino acid sequence identity (Figure 3-12B).

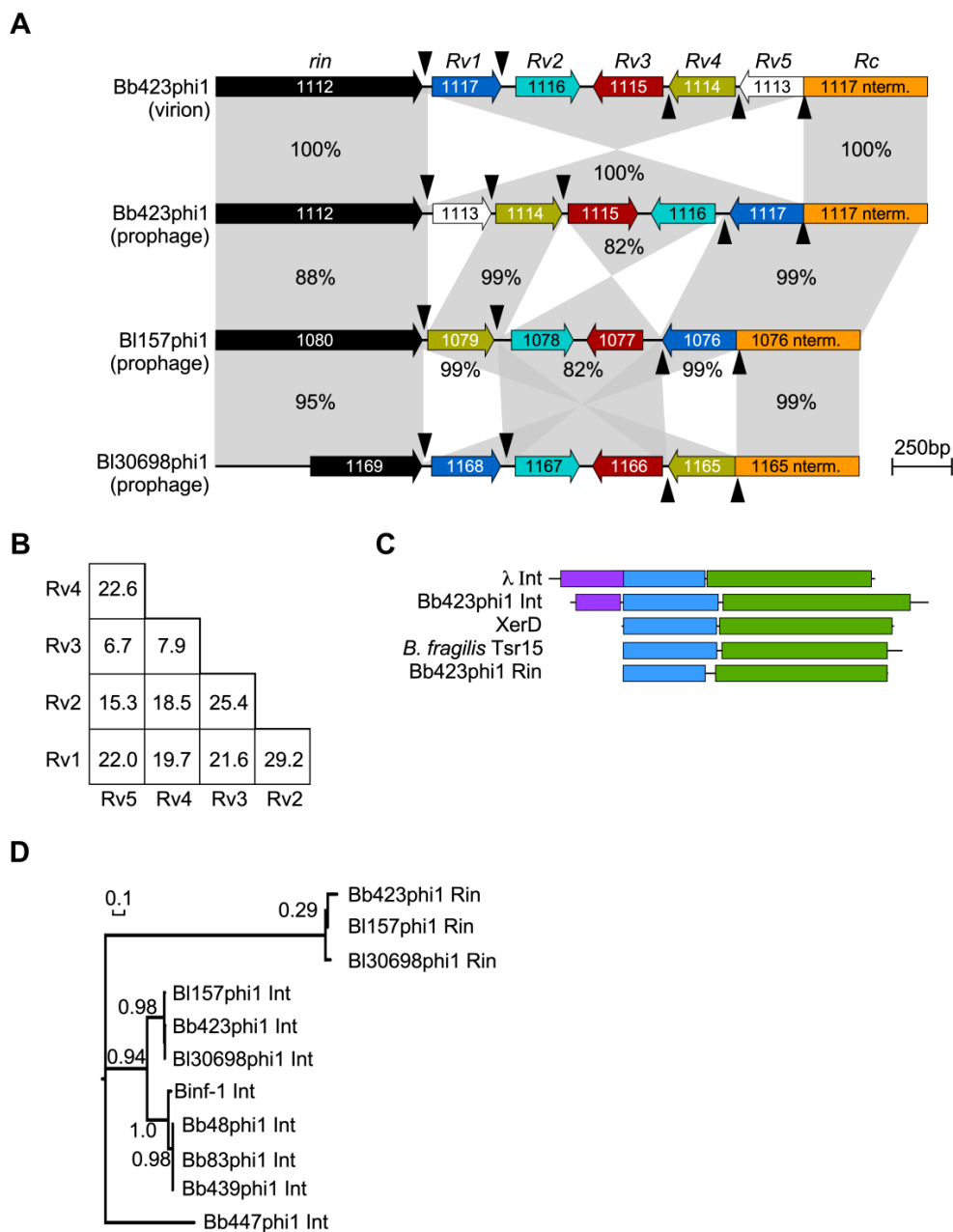


Figure 3-12. Characterization of the Rin shufflon.

(A) Enlarged view of the Rin shufflon from Figure 3-11, with genes (arrows) oriented relative to direction of transcription and labeled with their systematic gene numbers. BLAST alignment of the Rin shufflon loci from the Bb423phi1 prophage and induced virion genomes and BI157phi1 and BI30698phi1 prophage genomes highlight multiple sequence inversions. Shaded regions between genomes indicate regions of homology and percent sequence identities are labeled. Rin shufflon components analogous to the Min shufflon are indicated. Variable RBP 3' end coding regions (*Rv*) are numbered according to orientation in Bb423phi1 prophage and color-coded to highlight homologs in the BI157phi1 and BI30698phi1 prophages. *Rv* genes are flanked by the predicted tyrosine invertase

(*rin*, black) and the constant RBP 5' end coding sequence (*Rc*, orange). *Rv* genes are separated by putative 11 bp crossover sites (*rix*, arrowheads). **(B)** Matrix of pairwise Bb423phi1 *Rv* amino acid sequence identities. **(C)** Protein domain comparison between different types of tyrosine recombinases, including λ integrase (Int), the predicted Bb423phi1 integrase, XerD, *Bacteroides fragilis* Tsr15 invertase, and Bb423phi1 Rin. Approximate regions of the arm-type DNA-binding (purple), common core- DNA-binding (blue), and catalytic (green) domains predicted by HHpred are indicated. Proteins are manually aligned by the N-terminus of the common core DNA-binding domain. **(D)** Unrooted maximum likelihood phylogenetic tree constructed from alignment of the invertases and integrases identified in *B. breve* and *B. longum* prophages, with aLRT branch support indicated. Figure adapted from (Mavrich et al., 2018).

Several aspects of the Rin system are not found in other phase variation loci. Canonical invertases are related to the serine family of recombinases (Johnson, 2015). In contrast, Rin is predicted to be a member of the tyrosine recombinase family (Figure 3-12C). Structurally, it resembles previously characterized tyrosine recombinases such as XerD (Subramanya et al., 1997) or the Tsr15 gene associated with sequence inversion in *Bacteroides fragilis* (Weinacht et al., 2004). These tyrosine recombinases contain a common core-type DNA-binding domain and a catalytic domain, but they lack the N-terminal arm-type DNA-binding domain required for sequence-specific integration present in the tyrosine integrase of λ , or the predicted tyrosine integrases of these bifidophages. Furthermore, the Rin alleles are not closely related to any of the predicted bifidophage tyrosine integrases (Figure 3-12D). The absence of the arm-type DNA-binding domain suggests Rin is capable of directionless recombination like XerD, Tsr15, or the Integrase of *Mycobacterium* phage Brujita (Lunt and Hatfull, 2016), as would be required for facilitating sequence inversions at this locus. Lastly, the predicted *rix* sites are only 11 bp long and are asymmetric, unlike the longer inverted sequences utilized by the serine invertases (Johnson, 2015).

The Rin system appears to be active. The three homologous prophage loci exhibit high sequence similarity and several apparent inversions (Figure 3-12A). Multiple inversions can also be detected during assembly of the Bb423phi1 virion genome from the mitomycin C-treated lysate. In this sequencing sample, three contigs from the Bb423phi1 genome are assembled (Figure 3-13A). However, they are unable to be unambiguously combined into the complete genome due to sequencing reads that span different sets of contigs (Figure 3-13B, C). The sequencing data are consistent with a mixed population of phages exhibiting three variant Rin orientations due to two consecutive inversions (Figure 3-13D), although inversion of the entire segment is the predominant orientation (Figures 3-12A, 3-13B). Three orientations can also be detected during assembly of the *B. breve* 139W4-23 host genome (Figure 3-14, analysis performed by Francesca Bottacini). Altogether, five unique variant orientations of the Rin locus can be detected, in which the inversions occur precisely at the identified *rix* sites, and in which all possible *Rc-Rv* hybrid genes are identified.

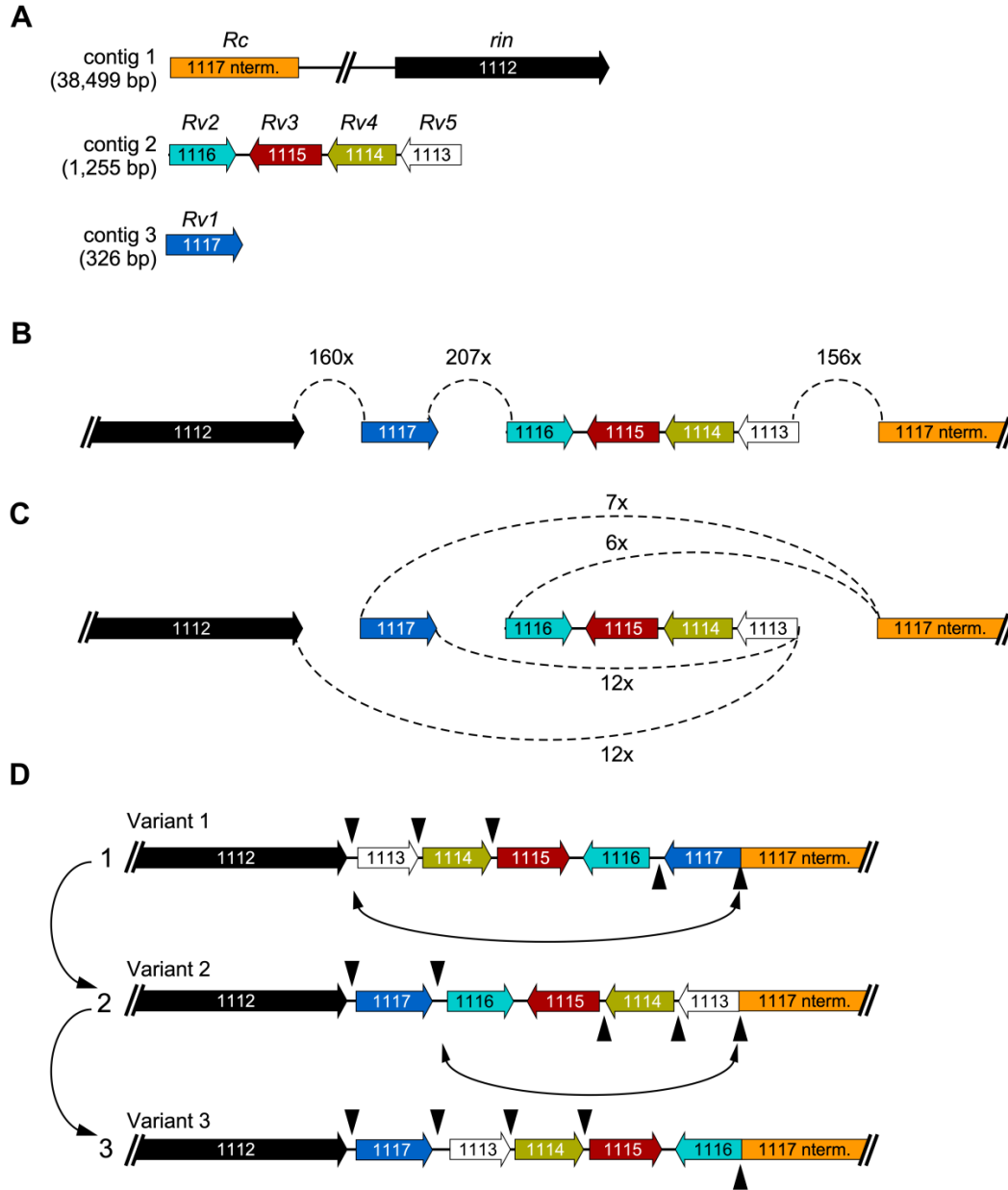


Figure 3-13. Induced Bb423phi1 virion genome harbors multiple Rin shufflon variants.

(A) Three contigs (numbered by size) representing the entire phage genome are assembled by Newbler, but a 100% consensus of the complete genome is not achieved. Contigs can be connected in multiple arrangements due to reads mapping across more than one contig, and these discrepant reads occur near or within the RBP locus (colors and gene numbering as in Figure 3-12A). (B) One possible contig orientation involves reads that straddle the three contigs (dashed lines) with approximately equal coverage. (C) Other contig orientations are possible, but they are represented by much lower read coverage and they do not obviously assemble into a single alternative genome. (D) Two sequential inversion events (double arrows) at *rix* sites result in three shufflon variants that sufficiently account for all hybrid sequence reads. Figure adapted from (Mavrich et al., 2018).

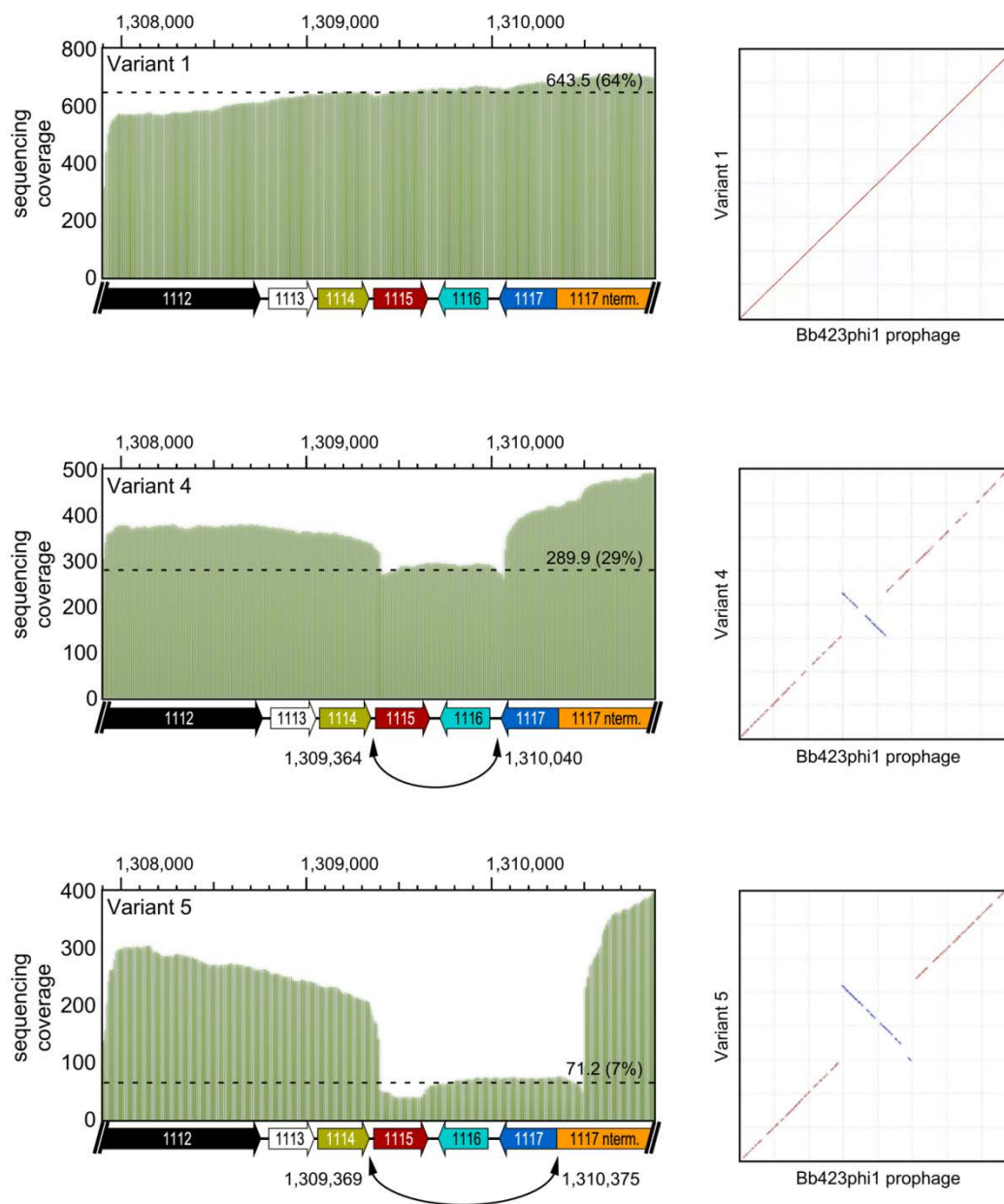


Figure 3-14. Uninduced Bb423phi1 prophage genome harbors multiple Rin shufflon variants.

Analysis of the previously reported *B. breve* 139W4-23 genome sequencing reads (Bottacini et al., 2017) reveals three variant orientations of the Bb423phi1 prophage Rin shufflon. (left) The variant nucleotide sequence orientations were assembled and all reads in the sample were mapped to each variant. The genome map below the histogram and the coordinates above the histogram reflect the predominant variant in the published genome. The points of inversion in each variant are indicated by the arrows, and the percentage of all reads in the sample that map to the variant orientation is indicated (colors and gene numbering as in Figure 3-12A). (right) Dot plot sequence comparison of each variant to the published prophage locus orientation highlights the points of inversion. Note: data analysis and figure were generated by Francesca Bottacini. Figure adapted from (Mavrich et al., 2018).

3.3.8 tRNA^{Met}-integrated prophages harbor an inversion locus

A second putative phase variation system is present in the two tRNA^{Met}-integrated prophages, 689b-1 and Bb423phi2 (Figure 3-15A). A ~ 500 bp sequence in Bb423phi2 containing a small gene, *BB139W423_0332*, and the 3' end of the adjacent gene, *BB139W423_0333*, have become inverted relative to 689b-1 (Figure 3-15B). Neither gene has an identifiable function based on homology searches. Similar to the Rin system, the observed inversion occurs at an 8 bp sequence, CAGGGTTA, and the two 3' end segments are quite dissimilar. However, unlike Rin, no recombinase is adjacent to the locus. Some invertases, such as the *Bacteroides fragilis* Mpi serine invertase, can act globally to facilitate inversions in *trans* (Coyne et al., 2003). Thus, this second bifidophages inverted locus may be a simpler phase variation system that relies on a DNA invertase supplied in *trans* by the host, which has not been previously reported in phage genomes.

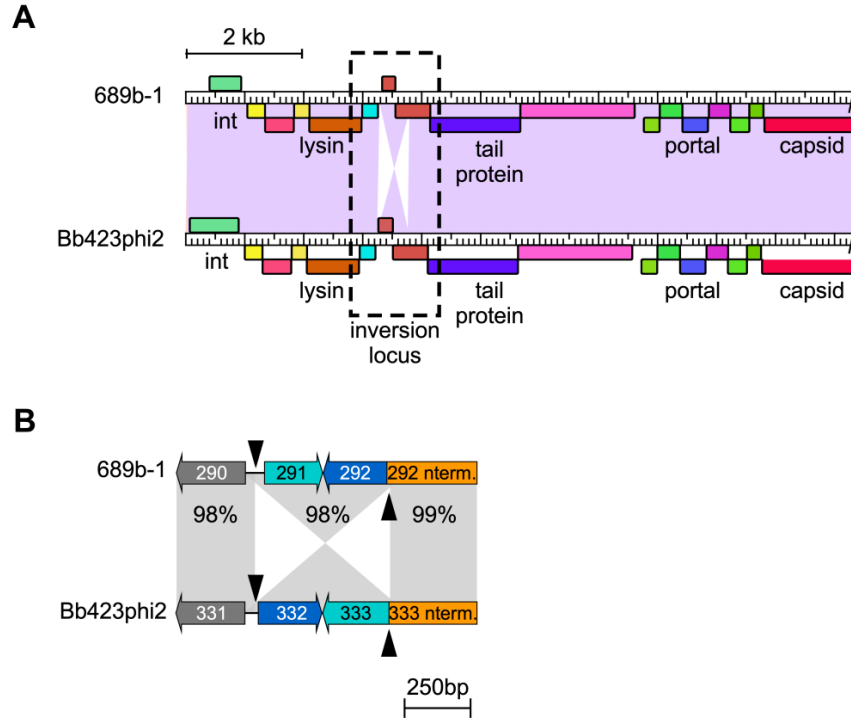


Figure 3-15. tRNA^{Met}-integrated prophages contain a phase variation system.

(A) Enlarged view of the left arm genes of tRNA^{Met}-integrated phages from Figure 3-1C highlights a putative phase variation system in these genomes. (B) Enlarged view from panel A of the inverted locus in tRNA^{Met}-integrated prophages, labeled as in Figure 3-12A. Figure adapted from (Mavrich et al., 2018).

3.3.9 Analysis of bifidophage host ranges

Comparison of the bifidophage integration sites also highlights interesting host range dynamics. Tyrosine integrases, as in the *Mycobacterium* phage L5 (Lee et al., 1991), utilize a homologous “common core” sequence present in both the *attP* and *attB* sites to facilitate integration. Although strand exchange occurs only within a 7-8 bp segment within the attachment sites, substantial sequence similarity is typically present across the entire common core sequence. However, unlike the *attL* and *attR* sites of the two tmRNA-integrated prophages or of B1157phi1, most of the *dnaJ*₂-integrated prophages do not have exactly matching attachment sites [Table 3-4, adapted from (Mavrich et al., 2018)].

Table 3-4. Bifidoprophage *attL* and *attR* common core sites.

Integration locus	Prophage	<i>Att</i> site	Sequence
<i>dnaJ₂</i>	Bb48phi1	<i>attL</i>	TTCTTTAGCAAGTTAAAAGACGCACTGAGCTGAGA
		<i>attR</i>	TTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGATT
	Bb83phi1	<i>attL</i>	TTCTTTAGCAAGTTGAAGGACGCACTGAGCTGAGA
		<i>attR</i>	TTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGATT
	Bb423phi1	<i>attL</i>	TTCTTTAGCAAGTTGAAGGACGCGCTGAGCTGAGA
		<i>attR</i>	TTCTTCGCAAGTTAAAGGATGCCTTGAGCTGATT
	Bb439phi1	<i>attL</i>	TTCTTTAGCAAGTTGAAGGACGCACTGAGCTGAGA
		<i>attR</i>	TTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGATT
	Binf-1	<i>attL</i>	TTCTTCAGCAAGTTAAAAGACGCACTGAGCTGAGA
		<i>attR</i>	TTCTTCTGCAAGTTAAAAGACGCACTGAGCTGAGA
	Bl30698phi1	<i>attL</i>	TTCTTCAGCAAGTTGAAGGACGCGCTGAGCTGAGA
		<i>attR</i>	TTCTTCGCAAGTTGAAGGACGCACTGAGCTGAGA
tmRNA	Bb447phi1/ Bb1192phi1	<i>attL/attR</i>	GTGGAGTCGCGGGGAATCGAACCCCG
tRNA ^{Met}	Bb423phi2/ 689b-1	<i>attL</i>	TGGTAGCGGGGCATGGATTTGAACCTTGGACCTCTGGGT
		<i>attR</i>	TGGTAGCGGGGCATGGATTTGAACCATGGACCTCTGGGT

Alignment of the *attL* and *attR* for the seven *dnaJ₂*-integrated prophages, including the *attP* of the four induced virion genomes, indicate the likely point of strand exchange and the flanking sequence differences (Figure 3-16A). Interestingly, the *attP* sites present in the induced *B. breve* phages are more similar to the *attB* sites present in *B. longum* genomes than to *attB* sites present in *B. breve* genomes (Figure 3-16B). For instance, the *attP* site in Bb48phi1 is more similar to the reconstructed *attB* site in *B. longum infantis* ATCC 15697 used by Binf-1 than to the reconstructed *attB* site in *B. breve* 082W4-8 used by Bb48phi1, and the *attP* site in Bb439phi1 is more similar to the reconstructed *attB* site in *B. longum longum* 157F used by Bl157phi1 than the reconstructed *attB* site in *B. breve* 017W4-39 used by Bb439phi1. This suggests that the four *B. breve dnaJ₂*-integrating phages may be able to integrate into multiple bifidobacterial species, or that there is common exchange of integration modules between bifidophages.

A

		strand exchange
Bb48phi1	<i>attL</i>	AGAAAGGG- - ATTCTTTAGCAAGTTAAAAGACGCACTGAGCTGAGAACGGG
	<i>attR</i>	CGAATGGTGGTTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGATTCGTTTC
	<i>attP</i>	CGAATGGTGGTTCTTCAGCAAGTTAAAAGACGCACTGAGCTGAGAACGGG
Bb83phi1	<i>attL</i>	AGAAAGGG- - ATTCTTTAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
	<i>attR</i>	CGAATGGTGGTTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGA- - - - -
	<i>attP</i>	CGAATGGTGGTTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
Bb423phi1	<i>attL</i>	AGAAAGGG- - ATTCTTTAGCAAGTTGAAGGACGCGCTGAGCTGAGAACGGG
	<i>attR</i>	CGAATGGTGGTTCTTCGCAAGTTAAAGGATGCCTTGAGCTGATTCGTTTC
	<i>attP</i>	CGAATGGTGGTTCTTCGCAAGTTGAAGGACGCGCTGAGCTGAGAACGGG
Bb439phi1	<i>attL</i>	AGAAAGGG- - ATTCTTTAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
	<i>attR</i>	CGAATGGTGGTTCTTCAGCAAGTTAAAGGATGCCTTGAGCTGA- - - - -
	<i>attP</i>	CGAATGGTGGTTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
BI157phi1	<i>attL</i>	AGAAAGGT- - TTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
	<i>attR</i>	CGAATGGTGGTTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGCGGA
BI30698phi1	<i>attL</i>	AGAAAGGT- - TTCTTCAGCAAGTTGAAGGACGCGCTGAGCTGAGAACGGG
	<i>attR</i>	CGAATGGTGGTTCTTCGCAAGTTGAAGGACGCACTGAGCTGAGAGCGGA
Binf-1	<i>attL</i>	AAAGAAAGGCTTCTTCAGCAAGTTAAAAGACGCACTGAGCTGAGAACGGG
	<i>attR</i>	TGAATGGTGGTTCTTCGCAAGTTAAAAGACGCACTGAGCTGAGAGTGGG

B

		strand exchange
Bb48phi1	<i>attB</i>	F F S K L K D A L S *
	<i>attP</i>	GAAGAAGGGATTCTTTAGCAAGTTAAAGGATGCCTTGAGCTGA- - - - -
Binf-1	<i>attB</i>	AAAGAAAGGCTTCTTCAGCAAGTTAAAAGACGCACTGAGCTGAGAGTGGG
		F F S K L K D A L S *
Bb439phi1	<i>attB</i>	F F S K L K D A L S *
	<i>attP</i>	AGAAAGGGA- - TTCTTTAGCAAGTTAAAGGATGCCTTGAGCTGA- - - - -
BI157phi1	<i>attB</i>	CGAATGGTGGTTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGTGGG
		AGAAAGGT- - TTCTTCAGCAAGTTGAAGGACGCACTGAGCTGAGAGCGGA
		F F S K L K D A L S *

Figure 3-16. *dnaJ*₂-integrated bifidoprophage attachment site analysis.

(A) For each *dnaJ*₂-integrated phage, the *attL* and *attR* sites, as well as their *attP* sites (if they were induced and sequenced), were aligned to determine the site of strand exchange during integration and excision and to deduce the *attB* sequence. Variant nucleotide positions are highlighted (beige). (B) The *attP* sites of Bb48phi1 and Bb439phi1 are aligned to the *attB* sites of their originating *B. breve* hosts and *B. longum attB* sites used by phages Binf-1 and BI157phi1. Figure adapted from (Mavrich et al., 2018).

3.3.10 Analysis of bifidophage evolutionary modes

The analyses in Chapter 2 revealed that there are two classes of temperate phages, marked by distinctly different rates of gene content flux (GCF). In general, the seven *dnaJ₂*-integrating prophages exhibit high GCF characteristics, and several of them can be classified as Class 1 (Figure 3-17). Thus, this group of bifidophages may experience more frequent horizontal gene exchange than other Class 2 temperate phages. Phages infecting the host phylum Bacteroidetes are commonly associated with the gut environment, and they exhibit similar evolutionary patterns, suggesting host environment may impact phage evolutionary modes (Figure 2-14B).

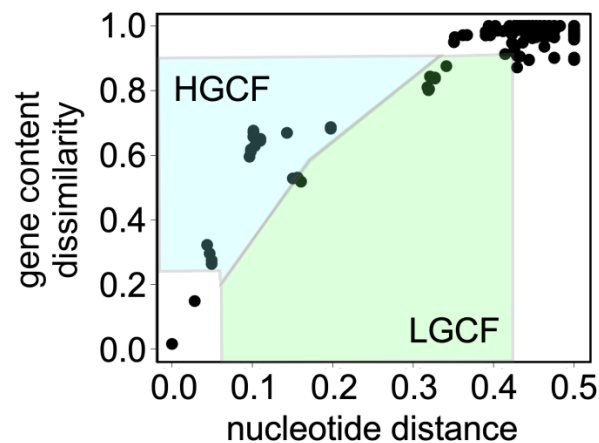


Figure 3-17. *dnaJ₂*-integrated prophages exhibit high gene content flux.

Pairwise comparisons (black circles) of nucleotide sequence and gene content between *dnaJ₂*-integrated phages and all other actinobacteriophages as previously described (Chapter 2) to highlight gene content flux patterns, with high (HGCF, blue) and low (LGCF, green) gene content flux regions indicated. Figure adapted from (Mavrich et al., 2018).

3.4 DISCUSSION

I have identified several bifidophages integrated at either the *dnaJ₂*, tmRNA, or tRNA^{Met} loci. Similar to other reported bifidophages, no plaques were observed despite repeated attempts using multiple strategies. Nevertheless, it is likely that they are currently, or are recent derivatives of, phages that are capable of forming infectious particles. They contain many of the genes required for all stages of phage growth and replication. Excision, replication, packaging, assembly, and lysis is observed (although to varying degrees) following mitomycin C treatment.

Characterization of these bifidophages can enhance our understanding of both actinobacteriophage and bifidobacterial biology. First, the bifidophages exhibit only distant similarity to other isolated and sequenced actinobacteriophages, highlighting that phages infecting *Bifidobacterium* hosts likely harbor substantial unexplored genetic diversity. Second, the integration of temperate bifidophages at *dnaJ₂* is particularly interesting. The most common types of genes that temperate phages utilize for site-specific integration are tRNA genes. This is likely beneficial since these genes are highly conserved and thus provide a reliable integration site. Other types of RNA genes, such as tmRNA (Pope et al., 2011b), and coding genes, such as *groEL* (Kim et al., 2003), are less commonly identified. DnaJ₂ is a highly conserved molecular chaperone that is present only in Actinobacteria, and it may play a distinct physiological role compared to its homolog DnaJ₁, which is found in all bacterial phyla (Ventura et al., 2005a). Thus, utilization of *dnaJ₂* for integration is an Actinobacteria-specific integration locus. Third, the multiple phase variation loci are notable. They are the first reported phase variation systems in actinobacteriophages. More broadly, Rin is the first reported phage-related system that utilizes a tyrosine recombinase, and the second phase variation loci is the first reported phage-related system that does not appear to utilize a recombinase in *cis*.

Technical aspects of studying bifidophages can be improved. First, the induction strategy could be improved. The effect of mitomycin C on induction metrics such as supernatant composition, excised phage copy number, and host growth inhibition is variable. Mitomycin C does not appear to be a robust strategy to induce these prophages. Flow cytometry could also be improved. The results only modestly reflect phage induction. This could be due to a combination of factors. Strains can have unique and unpredictable growth characteristics in different media. Most strains in this study grow well in RCM, so using this medium allowed for direct comparison between strains. However, in contrast to *Lactococcus* growth medium and other common bifidobacterial growth media, RCM produces much higher flow cytometric background signal which obscures the analysis. Additionally, spontaneous phage excision and release was observed in multiple strains, and low levels of phage present in untreated cultures would reduce mitomycin C-dependent fold changes and misleadingly suggest lower levels of phage are present.

Although the *dnaJ₂*-integrating bifidophages do not form plaques, they can nevertheless be used to further bifidobacterial research in several ways. First, the lysogens can be used as model strains to develop and improve bifidophage induction, isolation, and propagation strategies. Second, the strains can also be used as tools to study and manipulate natural gut microbial communities. Third, new bifidobacterial cloning vectors can be developed using the Actinobacteria-specific *dnaJ₂* integration module.

4.0 CHARACTERIZATION OF PARTITIONING SYSTEMS

Investigation of actinobacteriophage partitioning systems was a collaborative project published in the journal *Molecular Microbiology* (Dedrick et al., 2016). The following chapter primarily focuses on my contributions, but several experiments performed by others are included for clarity. Transcriptome profiling with RNAseq was performed in collaboration with Rebekah Dedrick and Dan Russell. Measuring RedRock copy number with DNAseq was performed by Rachael Rush, Rebekah Dedrick, and Dan Russell. RedRock and Alma ParB purification and EMSAs were performed by Juan Cervantes Reyes, Wei Ng, and Rebekah Dedrick. Rachael Rush and I investigated the RedRock origin of replication. Matt Olm constructed plasmids to investigate how RedRock *parABS* stabilizes plasmids.

4.1 INTRODUCTION

Extrachromosomal enterobacteria phages, such as P1 and N15, have played fundamental roles in understanding how partitioning systems function and how they are utilized by prophages to maintain lysogeny. However, there are relatively few examples of characterized partitioning systems derived from prophages. Instead, the majority of characterized or predicted partitioning systems are derived from plasmids or bacterial chromosomes (Ebersbach and Gerdes, 2005; Gerdes et al., 2000; Livny et al., 2007; Petersen et al., 2009; Pinto et al., 2012; Schumacher, 2012; Wang et al., 2013). However, extrachromosomal prophages may be common. Recently, extrachromosomal phages have been reported in hosts spanning diverse genera, including

Staphylococcus (Utter et al., 2014), *Yersinia* (Hertwig et al., 2003), *Halomonas* (Mobberley et al., 2008), *Vibrio* (Hammerl et al., 2014; Zabala et al., 2009), and *Leptospira* (Bourhy et al., 2005; Zhu et al., 2015). Additionally, several phages infecting actinobacterial hosts have been reported to be extrachromosomal, such as *Streptomyces* phage pZL12 (Zhong et al., 2010) and *Mycobacterium* phages 40AC, CRB1, and RedRock (Hatfull, 2012; Stella et al., 2013), but they are not well characterized.

Characterization of extrachromosomal phages and the genetic strategies they utilize for inheritance can provide insight into temperate phage evolution, as well as improve development of new genetic tools to study their hosts. There are relatively few reported plasmids that can replicate in *M. smegmatis*, and these include pMF1 (Bachrach et al., 2000), pJAZ38 (Gavigan et al., 1997), and pAL5000 (Labidi et al., 1985). The most commonly used vectors are derivatives of pAL5000, which was originally isolated from *Mycobacterium fortuitum* (Labidi et al., 1985) and can be used in several mycobacterial species (Labidi et al., 1992). This plasmid is low copy number (5 copies per cell), and substantial work has been done to characterize it and to create high copy number derivatives (30-60 copies per cell) (Bourn et al., 2007; Labidi et al., 1992; Raney et al., 1990; Stolt and Stoker, 1996, 1997). The mycobacterial vectors vary in compatibility and host range, and although incompatible replicons can be strategically utilized, such as pAL5000 derivatives for *Mycobacterium* gene replacement (Pashley et al., 2003), expanding the repertoire of characterized partitioning systems may help to create additional compatible vectors.

Here we characterized 42 extrachromosomal actinobacteriophages and their predicted partitioning systems. There are three types of partitioning systems, and in contrast to the Type Ia systems present in enterobacteria phages P1 and N15, many of the actinobacterial partitioning

systems are Type Ib, consisting of *parA* and *parB* genes flanked by *parS* binding sites. RedRock and Alma ParB homologs bind to *parS*, and RedRock ParB is required for partitioning. Replicons carrying genetically related *parABS* cassettes exhibit incompatibility, in which co-inheritance of both replicons is destabilized. ParB is subject to weaker purifying selection than ParA, potentially enabling diversification of partitioning systems to avoid prophage incompatibility.

4.2 MATERIALS AND METHODS

4.2.1 Phamerator database construction

The database *Actinobacteriophage_706* was created using Phamerator (Cresawn et al., 2011), consisting of 706 actinobacteriophages and is available online (http://phamerator.webfactional.com/databases_Hatfull). Gene products are grouped into phamilies (“phams”) using kClust (Hauser et al., 2013).

4.2.2 Generation of phages and lysogens

All phages and lysogens used for experiments in this chapter were also used to study the Cluster A immunity system. Refer to Materials and Methods in Chapter 5 for a full description of how phages and lysogens were prepared.

4.2.3 RNAseq

Strand-specific transcription profiles were measured by isolating total RNA from *M. smegmatis* mc²155 or lysogen cultures in exponential growth using Middlebrook 7H9 medium, as well as 30 min and 2.5 h after infection of mc²155 with RedRock at a multiplicity of infection of three. DNA was removed using the DNA-free Kit (Ambion) and rRNA was depleted using the Ribo-Zero Kit (Illumina). Libraries were prepared using the TruSeq Stranded RNAseq Kit (Illumina) and run on an Illumina MiSeq: one lane for each RedRock sample and one multiplexed lane combining wild type *M. smegmatis* and several lysogens (Alma, Pioneer, and EagleEye). The fastq reads were analyzed for overall quality using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and trimmed at the 5' and 3' ends with Cutadapt (<https://cutadapt.readthedocs.org>) using the quality score option and a value of 30. Trimmed reads were mapped simultaneously to the *M. smegmatis* mc²155 and RedRock genomes with Bowtie2 (Langmead and Salzberg, 2012), all non-unique reads were discarded using the command line tool, *sed*, and the data was processed with SAMtools (Li et al., 2009) and BEDtools (Quinlan and Hall, 2010) to compute strand-specific coverage. Raw fastq data are deposited in the Gene Expression Omnibus (GEO) under accession GSE79010. Rebekah Dedrick prepared the RNA for sequencing, Dan Russell performed the sequencing, and I analyzed the data.

4.2.4 DNaseq

DNA from a 2 ml sample of the RedRock or L5 lysogen was extracted from late exponentially growing cells (OD₆₀₀ ~ 1.0) using the Wizard Kit (Promega) according to the

manufacturer's protocol. DNA was quantified using the Qubit and libraries were prepared using the TruSeq Library Kit (Illumina) according to the manufacturer's protocol. The libraries were run on an Illumina MiSeq and data was evaluated using CLC Genomics (CLC bio-Qiagen, Aarhus, Denmark). This experiment was performed, analyzed, and described by Rachael Rush, Rebekah Dedrick, and Dan Russell.

4.2.5 Phylogenetic analysis of partitioning cassettes

The *Actinobacteriophage_706* contains 42 phages with predicted partitioning cassettes [Supplementary Table 4-1, originally published in (Dedrick et al., 2016)]. The protein sequences for 41 other putative and characterized NTPase and centromere-binding protein (CBP) genes from partitioning cassettes were identified from the literature and retrieved from NCBI (Supplementary Table 4-1). This non-exhaustive list contains representative partitioning systems from each partitioning type (Ia, Ib, II, III, or unknown) based on how cassettes have been previously categorized (Ebersbach and Gerdes, 2005; Gerdes et al., 2000; Schumacher, 2012), from various replicon types (chromosomal, plasmid, or phage), and from various bacterial host genera. Some replicons contained a previously predicted NTPase with no predicted CBP. However, there is typically an ORF immediately downstream in an apparent operon with the NTPase, and this gene was used as a potential CBP in the phylogeny. Protein sequences were aligned with ClustalO in SeaView (Gouy et al., 2010) and a phylogeny was constructed using the BioNJ algorithm with observed distances. A bootstrap analysis was performed with 100 replicates. Trees generated using other methods were comparable. Phylogenies were visualized and appended with genomic data using Evolview (Zhang et al., 2012a).

4.2.6 Prediction of partitioning types

HHpred (Soding et al., 2005) was used to predict the types of partitioning system for the cassettes used in this study. First, each partitioning protein was analyzed using HHpred with the *pdb70_15Feb16* database and with default settings as of February 22, 2016. The top 100 domain hits per protein that exceeded a homologous relationship probability (as computed by the program) cutoff of 90% were retained. All structural domain hits identified for the group of partitioning cassettes that have been previously categorized as Type Ia, Ib, II, or III were assigned a partitioning type category as follows. If the domain was found in one or more proteins from only one partitioning cassette type category, the domain was assigned the same partitioning type category, reflecting that in this analysis the domain is only found in gene products of that particular partitioning type. If the domain was present in proteins from more than one partitioning type, it was categorized as “nonspecific”. Next, the frequency of each domain category was calculated for partitioning proteins in the entire set of 83 partitioning cassettes. Finally, stacked bar graphs reflecting domain frequencies were generated for each taxon to provide a qualitative measure of how similar each partitioning protein is to previously characterized partitioning proteins. The analysis was done separately for NTPase and CBP proteins.

4.2.7 ParA and ParB coevolution analysis

The rate of evolution of the actinobacteriophage *parA* and *parB* genes were analyzed as follows. Of the 42 phages in the *Actinobacteriophage_706* database, Echid, 40AC, and pZL12 were not used; their ParA and/or ParB proteins did not group with the rest of the

actinobacteriophages in the protein sequence phylogenies, suggesting they were too distantly related for meaningful comparison. Of the remaining 39 phages, redundant nucleotide sequences of each partitioning gene were removed, reducing the list to a total of 27 phages that contain unique *parA* and *parB* sequences available for analysis. The *parA* and *parB* genes were processed separately. Nucleotide sequences were aligned by codon using webPRANK (Loytynoja and Goldman, 2010) and processed using the *kaks* tool in the *seqinr* R package to compute the pairwise K_A , K_S , and K_A/K_S values. The K_A/K_S ratios for all pairwise comparisons that had a $K_S < 2.0$ were retained and a scatter plot of the matching *parA* and *parB* ratios was generated.

4.2.8 Prediction of *parS-L* and *parS-R* loci

To systematically predict *parS-L* and *parS-R* sites, ~ 550 bp sequences straddling the start codon of the *parA* gene and the stop codon of the *parB* gene in each genome were individually analyzed for tandem sequence repeats using *etandem* on the EMBOSS server (Rice et al., 2000) as of January 14, 2016 with a defined potential repeat sequence length of 4-18 bp, and otherwise using the default settings. This program identifies regions in the query sequence that contain tandem repeats, and it computes the consensus repeat sequence, its frequency, and the % identity of all repeats in the repeat region. The consensus repeat sequences were aligned in CLC Genomics Workbench 8.5.1 (CLC bio-Qiagen, Aarhus, Denmark). Similar patterns were observed using Tandem Repeats Finder (Benson, 1999) or manual assessment with Gepard dot plot analysis (Krumsiek et al., 2007). Default parameters were used to identify tandem repeats, which may produce variations in repeat consensus sequence, length and frequency that are not biologically relevant. Although gross differences between some genomes are apparent and

noteworthy, more minor differences require closer inspection. For instance, after alignment of the consensus sequences, it is clear that some repeats that appear different are simply out of register by a few nucleotides, potentially due to minor sequence variations. The analysis was used to inform manual curation of *parS* loci for RedRock, Gladiator, Alma, Echid, and KatherineG by Graham Hatfull.

4.2.9 ParB Purification

The *parB* gene was PCR amplified from RedRock or Alma and inserted into plasmid pET-28a (Novagen) such as to include a C-terminal His tag to create pJC02 and pWN01, respectively (Appendix C). After verification by sequencing, the plasmids were transformed into BL21*(DE3)pLysS cells and grown to an $OD_{600} \sim 0.5$ at 37°C in LB medium. ParB-His expression was induced by the addition of 1 mM IPTG at 37°C for 3 h. Cells were pelleted, resuspended in 5 ml/g of Lysis Buffer (50 mM Tris pH 8.0, 300 mM NaCl, and 5% glycerol), and sonicated. The sonicated cells were centrifuged, and the cleared lysate was applied to a Ni-NTA column (Qiagen). The nickel column was washed with Lysis Buffer, 10 mM imidazole, and 50 mM imidazole, and the proteins were eluted with 150 mM imidazole. Fractions were collected and dialyzed in lysis buffer containing 30-50% glycerol overnight at 4°C. This experiment was performed and described by Juan Cervantes Reyes, Wei Ng, and Rebekah Dedrick.

4.2.10 Electrophoretic mobility shift assays

DNA substrates for electrophoretic mobility shift assays (EMSAs) were prepared using either gel-extracted PCR amplicons or annealed synthetic oligonucleotides (IDT and Invitrogen)[Table 4-1, adapted from (Dedrick et al., 2016)]. dsDNA substrates were radiolabeled at the 5' end with γ - ^{32}P using T4 polynucleotide kinase (Roche) at 37°C for 30 min and cleaned up using G-50 sephadex columns. EMSAs were performed by incubating 5-10 ng of radiolabeled substrates with serially diluted ParB at room temperature for 30 min in buffer (20 mM Tris pH 7.5, 10 mM EDTA, 25 mM NaCl, 10 mM spermidine, 1 mM DTT, and 1 μg calf thymus DNA) in a total volume of 10 μl . The DNA-protein samples were then resolved on a 5% native polyacrylamide gel run at 4°C. The gel was dried and exposed to a phosphorimaging plate, then scanned using a Fuji 5000 Phosphorimager. The dissociation constants (K_D) were calculated as the protein concentration at which 50% of the input DNA was bound by ParB. This experiment was performed and described by Wei Ng and Rebekah Dedrick.

Table 4-1. Description of ParB EMSA substrates.

Substrate	Phage Coordinates	Size (bp)
Nonspecific ^a	N/A	42
RedRock <i>parS-L</i>	27,232-27,897	666
RedRock <i>parS-R</i>	28,718-28,911	194
Gladiator <i>parS-L</i>	24,464-25,062	599
Gladiator <i>parS-R</i>	25,910-25,956	47
Alma <i>parS-L</i>	26,877-27,060	184
Alma <i>parS-R</i>	27,905-28,047	143
KatherineG <i>parS-L</i>	24,319-24,791	473
KatherineG <i>parS-R</i>	25,675-25,725	51
Echild <i>parS-L</i>	25,940-26,074	135

Note: substrates designed by Wei Ng and Rebekah Dedrick. ^aRefer to Appendix B for sequence.

4.2.11 Construction of *parABS* plasmids and plasmid retention assay

Plasmid pMO01 is a derivative of pLO87 (Oldfield and Hatfull, 2014), an extrachromosomal shuttle vector with P_{hsp60} upstream of *mCherry*, with the RedRock *parABS* cassette (coordinates 27,720-28,898) cloned downstream of *mCherry*. Several derivatives of pMO01 were constructed, including pMO02 and pMO03 (which contain a translational stop codon in the *parA* and *parB* genes, respectively) and pMO04 and pMO05 (which contain deletions of *parS-L* and *parS-R*, respectively)(Appendix C). Plasmids were transformed into *M. smegmatis* and grown in liquid culture with antibiotic selection for ~ 24 h or until saturation. Cultures were diluted 1:10,000 into liquid medium with no antibiotic selection and grown to saturation (~ 13 generations), and subsequent rounds of dilution were used to increase the number of rounds of unselected growth. Cultures were plated on solid medium, and colonies were scored for the presence (red) or absence (white) of the plasmid. Statistical significance of changes in plasmid retention were computed using two-sample two-tailed *t*-test of the retention level from three independent replicates. Matt Olm created these plasmids and performed the plasmid retention assay.

4.2.12 DNA skew analysis

DNA skews reflect strand-specific nucleotide biases across the sequence of interest. Cumulative AT and GC nucleotide skews were computed for genomes using GraphDNA (Thomas et al., 2007) using default settings.

4.2.13 Partitioning cassette incompatibility assay

Incompatibility of Cluster A partitioning cassettes was tested using a double lysogeny assay. In Stage 1, RedRock, Alma, Pioneer, and Bxb1 lysogens were generated by spotting phage lysates onto a confluent lawn of *M. smegmatis* mc²155, picking and clonally purifying cells from the center of the clearing, and verifying by PCR and standard superinfection immunity assays (see Chapter 5 Materials and Methods for more details). In Stage 2, 10⁷ PFUs of extrachromosomal (Pioneer or Alma) or integrating (Bxb1) phages from purified lysates were spotted onto a confluent lawn of the RedRock lysogen. After 2-3 days of growth at 37°C, cells from the center of the clearing were transferred to a 1.7 ml tube using a pipette tip, washed with Middlebrook 7H9 medium to minimize exogenous phage background, then spread onto Middlebrook 7H10 plates for single colonies. After 4-5 days of growth, single colonies were patched onto new Middlebrook 7H10 plates to again minimize exogenous phage background, allowed to grow for several days, and tested for double lysogeny with a spontaneous phage release assay on confluent lawns of wild type *M. smegmatis* mc²155, the original RedRock lysogen, and the lysogen of the challenging phage. Patches exhibiting phage release on both lysogens proceeded to Stage 3 where they were clonally purified three times on Middlebrook 7H10 plates, then tested a second time for double lysogeny as in Stage 2. In Stage 4 several of these purified strains that exhibited single or double lysogeny were grown in Middlebrook 7H9 liquid medium and tested for double lysogeny by a standard superinfection immunity assay as in Stage 1. Sample sizes of colonies treated and tested at each stage varied depending on the particular results of each phage pair. For Pioneer and Alma, two colony size morphologies were observed (normal and small), and colonies of each type were tracked through the experiment. Single phage infections of wild type *M. smegmatis* using each of the four phages were performed

in parallel with the double lysogen infections and carried through Stages 2 and 3 as controls in the spontaneous phage release tests.

4.2.14 Testing for a RedRock *parABS* origin of replication

To determine if the RedRock *parABS* is competent for replication, it was cloned into the vector pMD02, which does not contain an *M. smegmatis* origin of replication or integration cassette (Bibb and Hatfull, 2002). The locus spanning *parA* and *parB* genes and the flanking intergenic regions (coordinates 27,232-28,911) was PCR amplified using primers RR-16 (containing an *EcoRI* site) and RR-13 (containing a *BamHI* site), gel purified and cleaned up using the GeneJet Gel Extraction Kit. The amplicon and pMD02 were digested with *EcoRI* and *BamHI* restriction enzymes, and the pMD02 reaction was subsequently treated with calf intestinal phosphatase. Both restriction digestion reactions were gel purified as with the PCR reaction and ligated together to create pRR06. pRR06, pMD02, and a positive control vector containing the *oriM* cassette were transformed into electrocompetent *M. smegmatis* cells. Transformants were recovered for the positive control, but not for pMD02 and pRR06. Rachael Rush and I performed this experiment.

4.3 RESULTS

4.3.1 RedRock contains a partitioning cassette

There are hundreds of genetically related actinobacteriophages grouped together in Cluster A. Cluster A phages, such as the Subcluster A2 phage L5, exhibit a distinctive genomic architecture (Figure 4-1A)(Hatfull and Sarkis, 1993; Jain and Hatfull, 2000; Pope et al., 2011b). The left arm of the genome contains genes on the top strand and encode proteins associated with phage particle structure, assembly, packaging, and host lysis. The right arm consists of genes on the bottom strand and encode proteins associated with lytic growth such as DNA replication and DNA metabolism. Genes required for integration during lysogeny separate the two genome arms. They harbor a distinctive immunity system, in which the immunity repressor, positioned several kilobases downstream of the lytic promoter, P_{left} , regulates lysogeny utilizing 20-30 sites distributed throughout the genome (Figure 4-1A)(Brown et al., 1997; Ford et al., 1998; Jain and Hatfull, 2000; Pope et al., 2011b).

Several Cluster A phages, such as RedRock, represent an exception to this stereotypical architecture. Similar to L5, RedRock is grouped into Subcluster A2 (Figure 4-1A). These two phages exhibit substantial nucleotide sequence similarity across their genome, and the two phages are homoimmune (Pope et al., 2011b). However, in contrast to L5 and other Cluster A phages such as Bxb1 (Jain and Hatfull, 2000), D29 (Ford et al., 1998), and Peaches (Pope et al., 2011b), RedRock does not contain a predicted integrase gene. Instead, it contains a locus that exhibits similarity to replicon partitioning systems (Figure 4-1A, B).

(Soding et al., 2005) and the NCBI Conserved Domain Database (CDD). Gp38 is 89 amino acids and exhibits sequence similarity to several DNA-binding proteins related to partitioning and plasmid copy control, suggesting it acts as a ParB-like protein (Figure 4-1A, B). Additionally, 8 tandemly repeated 6-8 bp sequences can be identified immediately upstream of gene 37 (*parS-L*) and downstream of gene 38 (*parS-R*) that may be centromere-binding protein (CBP) binding sites (Figure 4-1B, C). The structure of the locus strongly suggests that RedRock is a temperate phage that utilizes a partitioning cassette to maintain lysogeny as an extrachromosomal prophage, similar to enterobacteria phage P1 (Austin and Abeles, 1983), instead of integrating into the genome.

Similar to RedRock, there are 40 other Cluster A phages infecting *Mycobacterium* and *Gordonia* hosts that lack an integrase and may replicate as extrachromosomal prophages, and I characterized their prophage inheritance genes (Supplementary Table 4-1). Similar to RedRock, they all contain tandem *parA*-related and *parB*-related genes at the center of their genomes (Figure 4-2A). The potential partitioning cassettes are structured very similarly, with *parA* and *parB* of similar sizes, and with *parA* upstream of *parB* (Figure 4-2A). A bioinformatic survey of the flanking intergenic regions indicates that most of these phages contain regions with short tandem sequence repeats adjacent to *parA* (~ 75-100 bp cumulative length) and adjacent to *parB* (~ 50-80 bp cumulative length), similar to *parS-L* and *parS-R* in RedRock (Figures 4-1, 4-2A, see Materials and Methods). The *parS* loci vary between phages, including the size and sequence of the consensus site, the number of tandem repeats, and the % identity of the repeats (Figure 4-2A, B). Only four phages, including Luchador (Subcluster A14), Loser (Subcluster A2), 40AC (Subcluster A2), and CRB1 (Subcluster A2) do not have predicted *parS-L* or *parS-R* loci (Figure 4-2A, B).

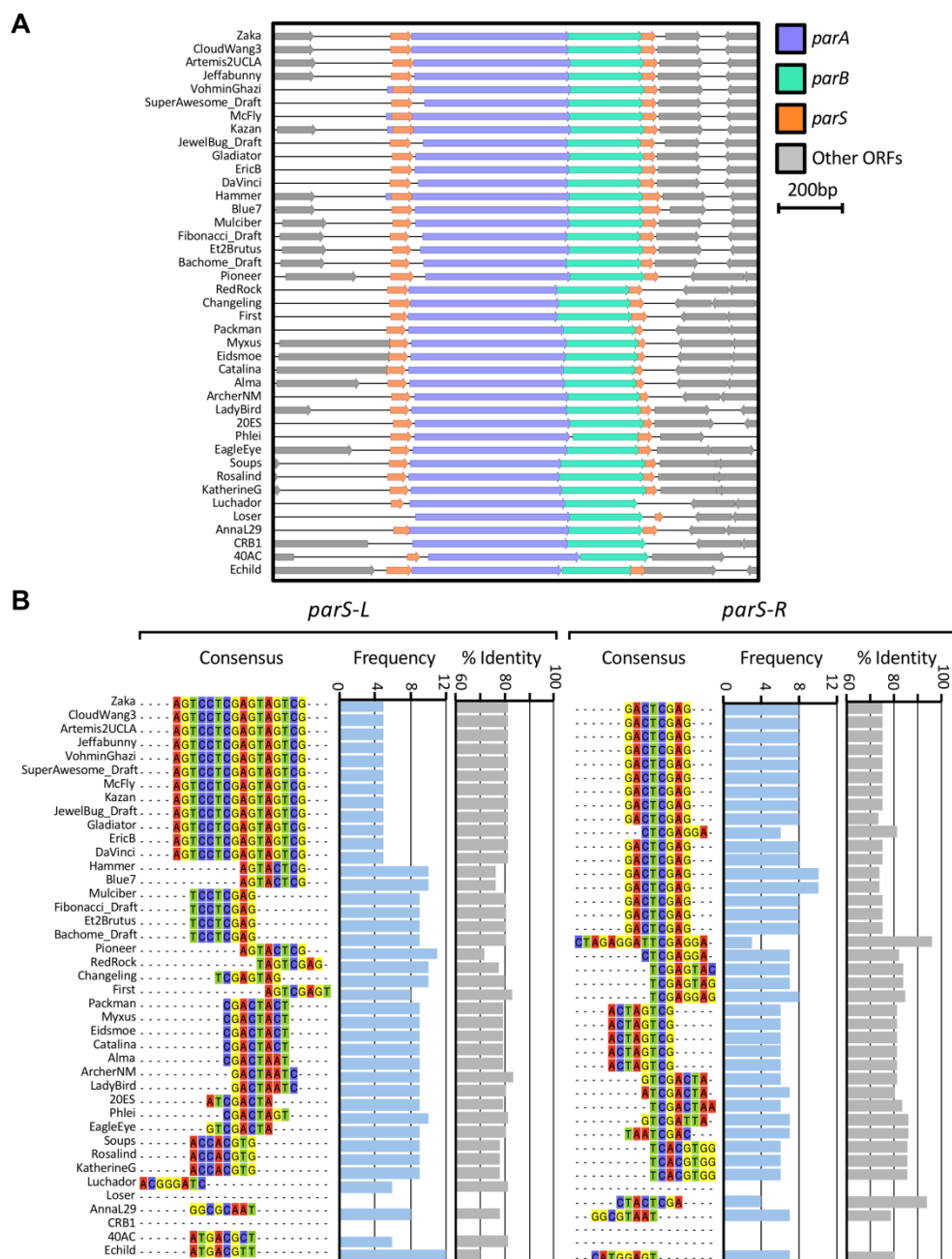


Figure 4-2. Characterization of predicted *parABS* systems.

(A) Enlarged view of partitioning systems from genome maps of 41 actinobacteriophages, with the predicted *parA*, *parB*, *parS-L*, and *parS-R* components indicated (Supplementary Table 4-1). (B) Alignment of the consensus repeat sequences predicted by *etandem* with loci ordered to match the ParB phylogeny in Figure 4-3B. The corresponding frequency and % identity of the repeats are plotted. Figure adapted from (Dedrick et al., 2016).

4.3.2 Characterization of Cluster A partitioning systems

Several types of partitioning systems have been characterized, and they exhibit distinct genetic architectures (Ebersbach and Gerdes, 2005; Gerdes et al., 2000; Schumacher, 2012). Additionally, partitioning systems have been predicted in phages isolated from hosts of other phyla (Supplementary Table 4-1, see Materials and Methods). The genetic relationships of the Cluster A partitioning systems to other characterized and uncharacterized systems are not clear. To characterize the genetic diversity of these prophage partitioning systems and to investigate how they may function, NTPase and CBP proteins from 83 predicted and characterized partitioning systems were compared, representing diverse partitioning systems derived from bacteria, phage, and plasmid replicons (Figure 4-3). The sizes, domains, and evolutionary histories of the NTPase and CBP proteins were compared (see Materials and Methods).

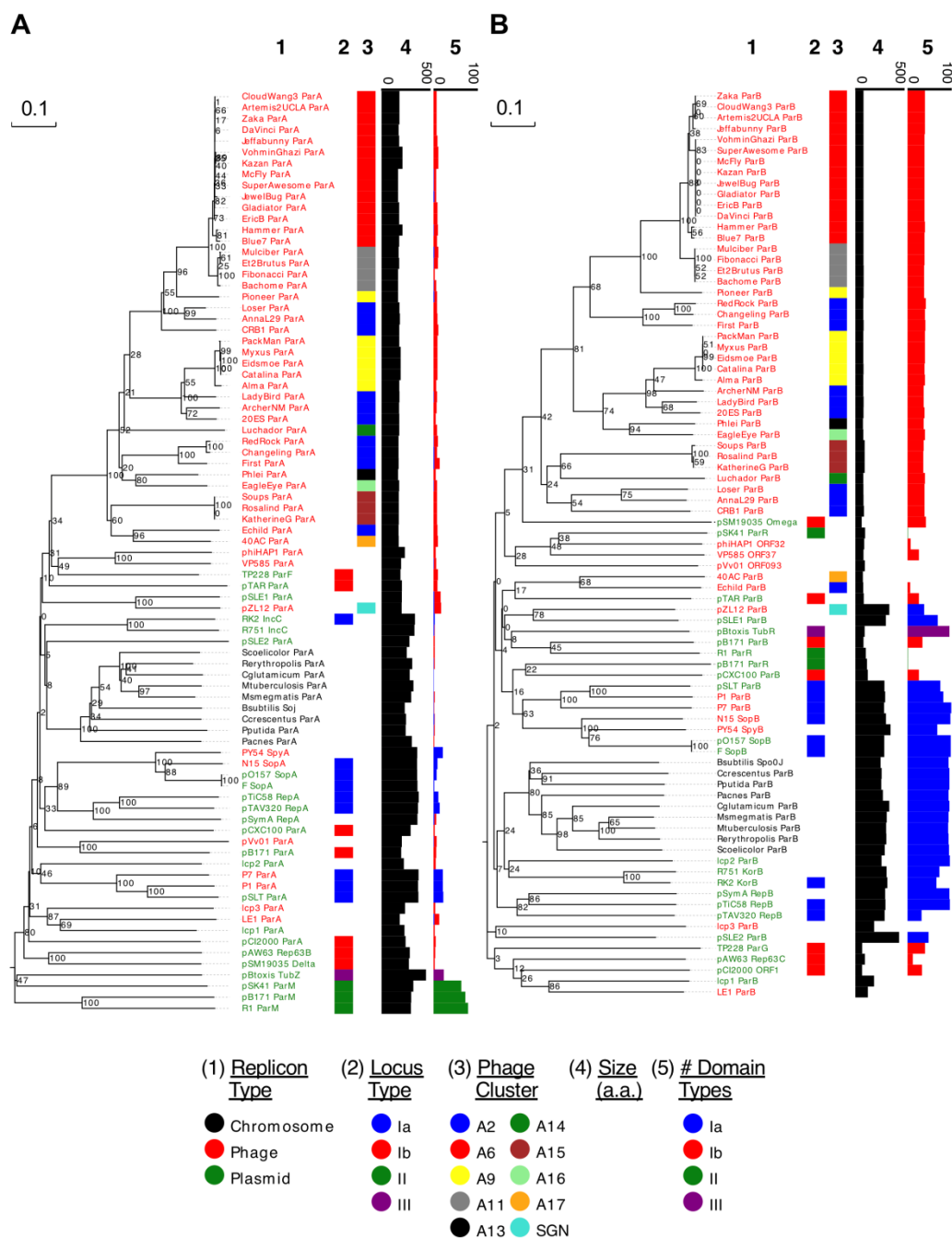


Figure 4-3. Phylogenetic comparison of NTPase and CBP proteins.

Phylogenetic trees were constructed for **(A)** NTPase and **(B)** CBP proteins from 83 characterized or predicted partitioning systems (1) derived from various replicons (chromosomal, plasmid, and phage) representing (2) various partitioning systems (Ia, Ib, II, and III). Branch supports from bootstrap analysis indicated. For each sequence, (3) subcluster designation is indicated if applicable, (4) the horizontal bar graph reflects amino acid size, and (5) the horizontal bar graph reflects frequency of categorized HHpred domains. Figure adapted from (Dedrick et al., 2016).

In general, the collection of NTPase and CBP proteins are diverse, as noted previously (Gerdes et al., 2000), such that phylogenetic branches are deep, and the overall structure of the trees at the deepest roots do not have strong branch supports (Figure 4-3A, B). However, all ParA proteins present in Cluster A actinobacteriophages form a monophyletic clade with strong branch support (Figure 4-3A), suggesting a common evolutionary origin. They are all ~ 180-230 amino acids, comparable to Type Ib NTPases and in contrast to the ~ 300-400 amino acid NTPases associated with Type Ia, II, and III systems (Supplementary Table 4-1)(Ebersbach and Gerdes, 2005). Additionally, the majority of partitioning-related domains present in these proteins are associated with Type Ib NTPases (Figure 4-3A). Similarly, Cluster A ParB proteins form a monophyletic clade, with the exception of the ParB proteins from Echid and 40AC (Figure 4-3B). Cluster A ParB proteins are all ~ 85-105 amino acids in size, similar to Type Ib CBPs and in contrast to the Type Ia, II, and III CBPs that range from ~ 100-350 amino acids (Supplementary Table 4-1)(Ebersbach and Gerdes, 2005). Additionally, the majority of partitioning-related domains present in these proteins are associated with Type Ib systems and are not closely related to Type Ia systems present in mycobacterial hosts (Figure 4-3B). Although the Cluster A partitioning systems do not appear to be closely related to other characterized systems, their ParA and ParB proteins are structured similarly to Type Ib systems.

The partitioning systems appear to be cassettes that can be exchanged between phages. For instance, Pioneer and Alma are very similar Subcluster A9 phages, exhibiting 97% nucleotide identity across 93% of their genomes (Figure 4-4). However, there is very little sequence similarity across their *parABS* loci, and Pioneer's partitioning genes do not phylogenetically group with the other Subcluster A9 partitioning proteins (Figure 4-3).

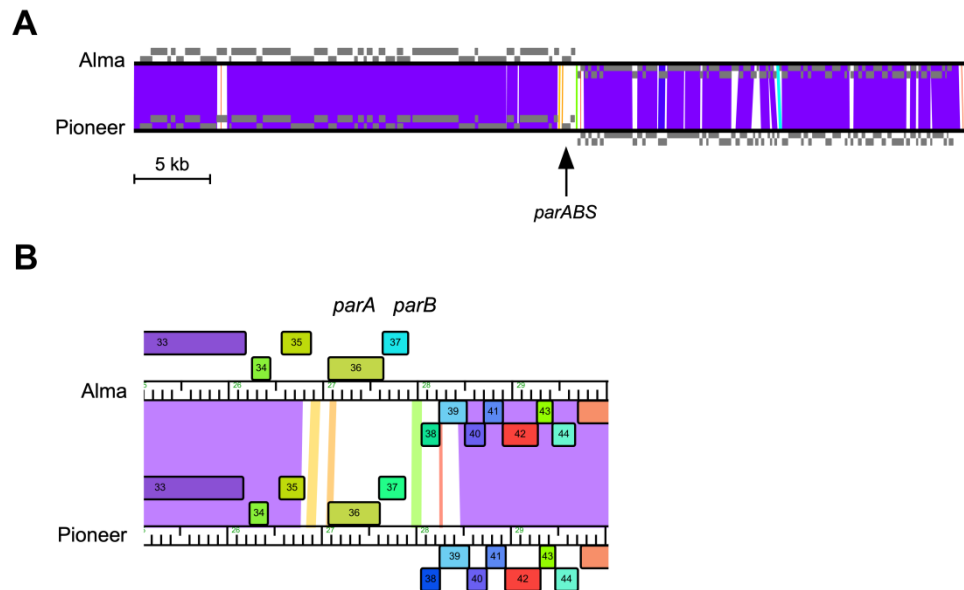


Figure 4-4. Subcluster A9 phages contain different *parABS* loci.

(A) Phamerator alignment of Pioneer and Alma. **(B)** Enlarged view of the *parABS* locus from panel A.

4.3.3 RedRock partitioning genes are expressed during lysogeny

Stable lysogens containing the RedRock prophage can be generated (Pope et al., 2011b). To determine when *parA* and *parB* are expressed relative to other genes, Rebekah Dedrick and I performed RNAseq on a RedRock lysogen and during several stages of RedRock lytic growth after infection of mc²155 (Figure 4-5A, B). During lysogeny, expression is observed at only a few loci across the RedRock genome. There is some expression at the right end of the genome near the putative lytic promoter, P_{left}, that has been mapped in L5 (Nesbit et al., 1995), as well as expression across the immunity repressor locus, as expected (Nesbit et al., 1995). Some expression is observed near the left end of the genome, although the biological impact of this locus is not clear. There is also expression at the *parABS* locus, in which expression begins upstream of *parA* and extends through *parB*. During early lytic growth (30 min post infection), increased expression is observed across the right arm of the genome, which encodes many genes associated with DNA metabolism and replication (Hatfull and Sarkis, 1993). At this stage of growth, there is also strong expression of *parA* and *parB*. During late lytic growth (150 min post infection), increased expression is observed across the left arm of the genome, which encodes many genes associated with phage particle structure and assembly (Hatfull and Sarkis, 1993). Relative to expression of these late genes, expression at *parABS* is very weak. Further work performed by Matt Olm shows that the region upstream of *parA* can drive expression of a reporter gene (Dedrick et al., 2016). Overall, the transcriptional profiles suggest *parA* and *parB* are expressed from an upstream promoter, P_{par}, during early lytic infection and lysogeny, consistent with the hypothesis that they are required to ensure prophage inheritance (Figure 4-1B).

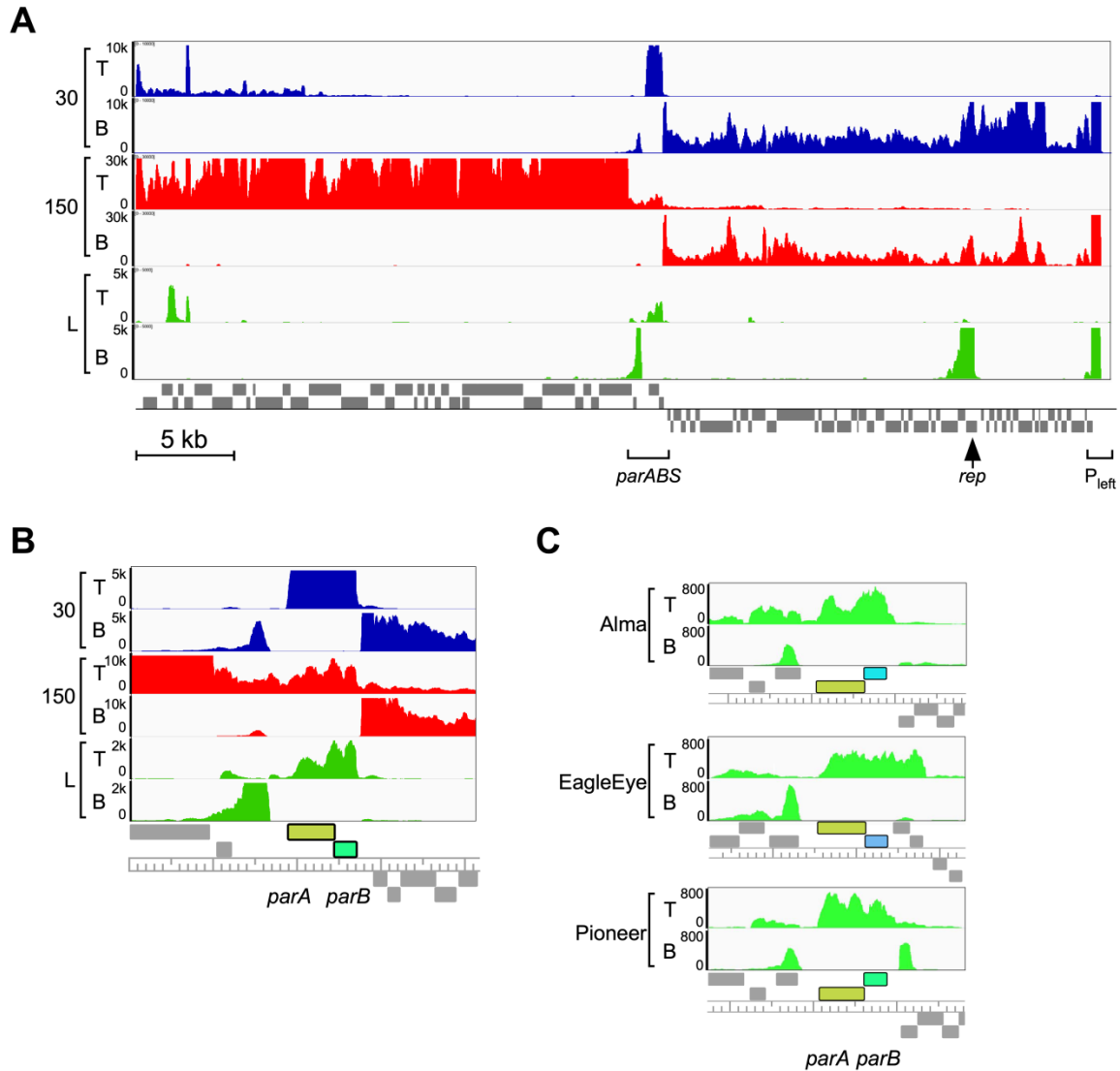


Figure 4-5. *parABS* systems are expressed during lysogeny.

(A) Strand-specific RNAseq expression profiles for top (T) and bottom (B) strands of RedRock (Subcluster A2) during lysogeny (L, green) and infection (30 min, blue, and 150 min, red), aligned to the RedRock genome map. (B) Enlarged view of the *parABS* locus from panel A. (C) Enlarged view of the *parABS* locus with strand-specific RNAseq profiles for Alma (Subcluster A9), Pioneer (Subcluster A9), and EagleEye (Subcluster A16) extrachromosomal phages during lysogeny (green). Rebekah Dedrick and I performed this experiment. Figure is modified from the original figure generated by Rebekah Dedrick (Dedrick et al., 2016).

Stable lysogens can be generated for many other Cluster A phages with *parABS* systems, including LadyBird and ArcherNM (Subcluster A2), Gladiator and DaVinci (Subcluster A6), Alma and Pioneer (Subcluster A9), Et2Brutus and Mulciber (Subcluster A11), and EagleEye (Subcluster A16)(although some prophages, such as ArcherNM and LadyBird, may be less stable than others, since liquid cultures of these lysogens appear to occasionally lose the prophage). Rebekah Dedrick and I performed RNAseq for some of these phages during lysogeny to compare expression profiles of *parABS* systems (Figure 4-5C). Alma, Pioneer, and EagleEye prophages all exhibit similar expression profiles across *parABS* as seen in RedRock. Therefore, the partitioning systems in these genetically diverse phages are likely performing similar functions during lysogeny.

4.3.4 Multiple copies of RedRock are maintained during lysogeny

Replicons that utilize partitioning systems are typically maintained at low copy number, such as prophage N15, maintained at 3-5 copies per cell (Ravin, 2015). In contrast, prophages that integrate into the host chromosome, such as L5, are expected to be maintained at one copy per cell. To determine the impact of *parABS* on copy number, Rachael Rush, Rebekah Dedrick, and Dan Russell compared nucleotide content in lysogens carrying an L5 or RedRock prophage. DNA was extracted from both lysogens during exponential growth and mapped to each lysogen genome sequence (see Materials and Methods). Sequencing coverage across the L5 prophage is comparable to coverage across the host genome (Figure 4-6). In contrast, sequencing coverage across the RedRock prophage is substantially higher compared to the host genome (Figure 4-6). Of the two sequencing reactions, approximately equal number of reads are mapped to the *M. smegmatis* genome, with an average coverage of ~ 62x (Table 4-2). However, coverage across

RedRock is more than twice as high as coverage across L5 (Table 4-2). Average coverage across L5 compared to the average coverage across the host approaches 1, consistent with L5 integrating into the host genome at a copy number of 1. In contrast, average coverage across RedRock compared to average coverage across the host suggests that RedRock is maintained at 2.41 copies per cell (Table 4-2). Furthermore, reads could be mapped to L5 virion genome ends and the L5 *attP* site, reflecting low levels of L5 spontaneous induction, but no reads could be mapped to RedRock virion genome ends (Table 4-2). These data are consistent with RedRock being an extrachromosomal prophage.

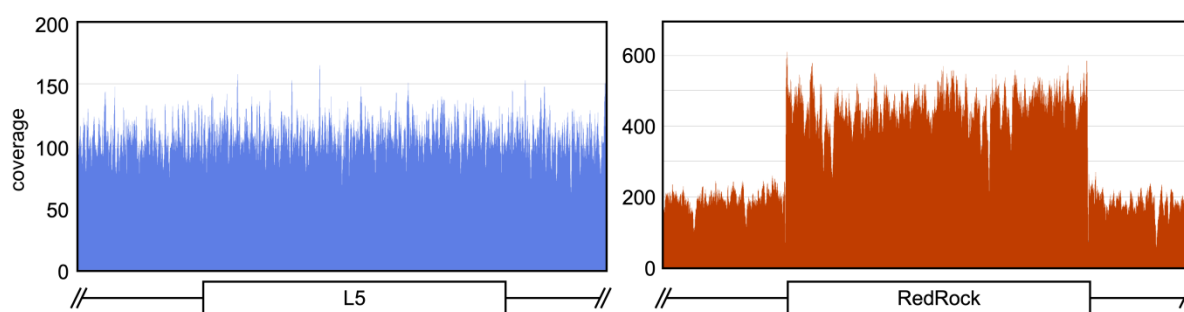


Figure 4-6. RedRock prophage exhibits increased copy number.

DNA was extracted from (left) L5 and (right) RedRock lysogen cultures and sequenced. Sequencing reads were mapped to the lysogen genomes in which both prophage genomes were integrated into the known L5 integration site. This experiment was performed by Rachael Rush, Rebekah Dedrick, and Dan Russell. Figure is modified from the original figure generated by Dan Russell (Dedrick et al., 2016).

Table 4-2. Comparison of RedRock and L5 prophage copy number.

Lysogen	Host coverage	Phage reads	Phage coverage	Phage:host coverage	Viral end reads	<i>attL/attR</i> reads	<i>attP</i> reads
L5	62.5	41,053	58.9	0.94	23	141	2
RedRock	61.8	105,978	149	2.41	0	N/A	N/A

Note: DNA sequencing, data analysis, and table presentation were performed by Rachael Rush, Rebekah Dedrick, and Dan Russell. Table adapted from (Dedrick et al., 2016).

4.3.5 RedRock ParB exhibits *parS* binding specificity

To determine whether Cluster A ParB homologs bind to the predicted *parS* sites, Juan Cervantes Reyes, Wei Ng, and Rebekah Dedrick purified the RedRock ParB and tested its binding specificity (Figure 4-7). RedRock ParB can bind to a 650 bp substrate that spans the intergenic region upstream of RedRock *parA* and contains *parS-L*, exhibiting a dissociation constant (K_D) \sim 100-300 nM. RedRock ParB exhibits some binding to a 200 bp substrate that spans the intergenic region downstream of *parB* and contains *parS-R*, although it produces less distinct bands and does not appear to bind as well. Binding is specific though, since RedRock ParB but not exhibit binding to a 42 bp sequence unrelated to *parS*. A more in-depth evaluation of RedRock ParB binding specificity has been performed by Wei Ng, showing that ParB binds weakly to a substrate that contains a single binding site, and exhibits complex binding behavior to substrates with additional sites (Dedrick et al., 2016), similar to binding behaviors of other Type Ib ParBs derived from plasmids pSM19035 (Dmowski et al., 2006) and TP228 (Carmelo et al., 2005; Zampini et al., 2009). Overall, RedRock ParB likely binds to sites within *parS-L*, and possibly *parS-R*, to promote prophage segregation. Consistent with the *in vitro* binding data, *in vivo* experiments performed by Matt Olm show that the promoter activity upstream of *parA* is regulated by ParB (Dedrick et al., 2016), as shown for other Type Ib systems (Gerdes et al., 2000; Schumacher, 2012).

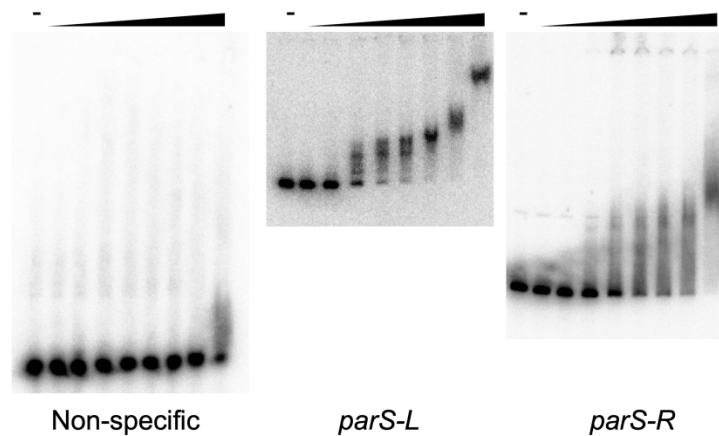


Figure 4-7. RedRock ParB exhibits *in vitro* binding affinity for *parS-L* and *parS-R*.

The binding affinity of purified RedRock ParB for different substrates was tested. Serially diluted ParB was incubated with radiolabeled PCR-amplified 666 bp *parS-L* and 194 bp *parS-R* substrates, as well as a 42 bp oligonucleotide substrate that is not predicted to contain any *parS* sites. This experiment was performed by Juan Cervantes Reyes, Wei Ng, and Rebekah Dedrick. Figure is modified from the original figure generated by Graham Hatfull (Dedrick et al., 2016).

4.3.6 RedRock *parABS* increases plasmid retention

Partitioning systems are expected to ensure efficient segregation of replicons to progeny during cellular growth and division. To test whether Cluster A partitioning systems also exhibit this characteristic, Matt Olm constructed a series of plasmids using *mCherry* as a reporter gene (Table 4-3). Expression of mCherry produces a red colony phenotype that can be used to measure plasmid retention. When it is expressed from the strong *hsp60* promoter in pLO87, the plasmid becomes destabilized, and in the absence of antibiotic selection, pLO87 is quickly lost (Oldfield and Hatfull, 2014). Matt Olm cloned different segments of the *parABS* locus into pLO87 and measured plasmid retention (Table 4-3). When the entire *parABS* locus is cloned, plasmid retention is substantially improved relative to pLO87, indicating the partitioning system

helps to ensure segregation of the plasmid during cellular growth. When *parS-R* is removed, plasmid retention remains quite high, indicating that although ParB is able to bind sites downstream of the operon, they are not essential for plasmid stability. However, when ParA is not translated, plasmid retention returns to levels comparable to pLO87, indicating that ParA is required for segregation. A plasmid carrying all *parABS* elements except for *parB* or *parS-L* could not be transformed into *M. smegmatis*, suggesting that expression of ParA alone is toxic to the cell. ParB may auto-regulate expression of the *parA-parB* operon using sites in *parS-L*, as reported for other Type Ib systems (Schumacher, 2012).

Table 4-3. RedRock ParB increases plasmid retention.

Plasmid	Reporter	<i>parABS</i> elements	Plasmid retention (%) ^a
pLO87	<i>hsp60-mCherry</i>	N/A	<1
pMO01	<i>hsp60-mCherry</i>	<i>parSL-parA-parB-parSR</i>	82 ^b
pMO02	<i>hsp60-mCherry</i>	<i>parSL-parA*-parB-parSR</i>	7
pMO05	<i>hsp60-mCherry</i>	<i>parSL-parA-parB</i>	71 ^b

^aPlasmid retention determined by percentages of red colonies after 52 generations of unselected growth.

^bDifference in pMO01 and pMO05 is not significant (p -value = 0.25). **parA* gene contains a stop codon that prevents complete translation of the open reading frame. Note: table is modified from the original table generated by Graham Hatfull (Dedrick et al., 2016).

4.3.7 RedRock *parABS* confers replicon incompatibility

Replicons that carry homologous partitioning systems exhibit incompatibility (Austin and Nordstrom, 1990; Ebersbach and Gerdes, 2005; Novick, 1987). This phenomenon may occur due to the inability for the NTPase and CBP to efficiently regulate the segregation of multiple replicons to progeny, resulting in the stochastic loss of one of the replicons. To determine whether the Cluster A *parABS* system exhibits a similar phenotype, I performed an incompatibility assay using the plasmid pMO01 and the RedRock prophage (Figure 4-8A).

Plasmid pMO01 is transformed into mc²155 or several lysogens, including RedRock, EagleEye (a Subcluster A16 phage that also carries a partitioning system), and L5 (which integrates into the host genome). Positive plasmid transformants are selected using kanamycin, but there is no selection for the prophage (see Materials and Methods). Positive plasmid transformants are cultured with selection and are assayed to determine if they have also retained the prophage by testing for spontaneous phage release. If the partitioning system exhibits incompatibility, the extrachromosomal prophage will be lost and no phage release should be observed.

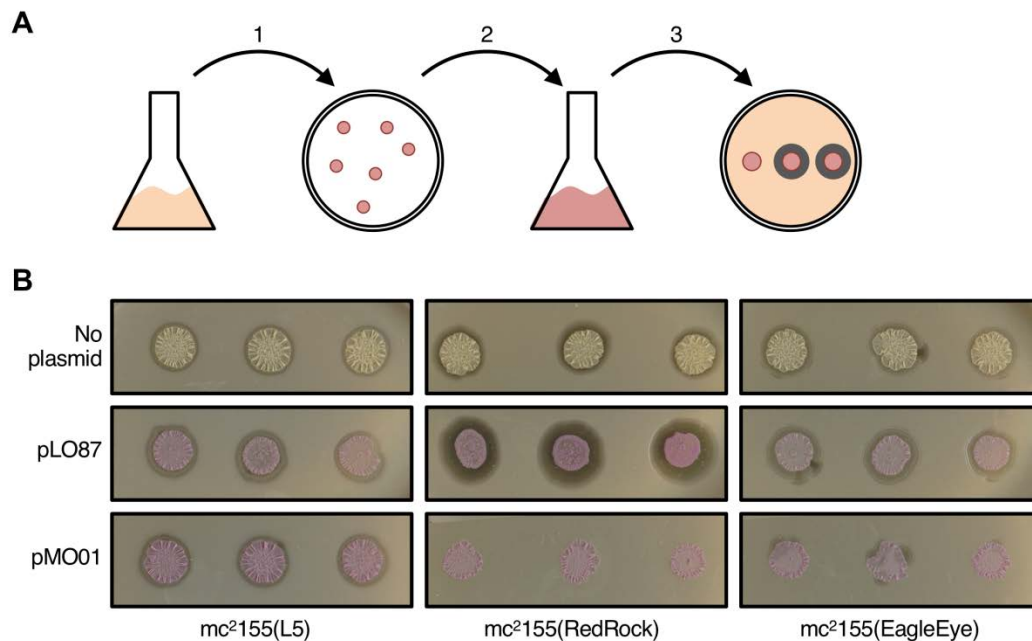


Figure 4-8. RedRock *parABS* promotes plasmid-prophage incompatibility.

(A) Experimental design to test for partitioning-mediated incompatibility between an extrachromosomal plasmid and prophage. (Step 1) Electrocompetent *M. smegmatis* lysogens (carrying prophages L5, RedRock, or EagleEye) were transformed with plasmids carrying *mCherry* with (pMO01) and without (pLO87) the RedRock *parABS* cassette, and positive transformants were selected. (Step 2) Three different positive transformants (or untransformed control colonies) were propagated in liquid selective media. (Step 3) Saturated cultures were spotted onto a lawn of *M. smegmatis* to test for phage release by identifying zones of inhibition. (B) Three independent transformants for each plasmid in each strain were spotted onto *M. smegmatis* mc²155 from Step 3 in panel A. Figure adapted from (Dedrick et al., 2016).

Selection for pMO01 results in a complete loss of the RedRock prophage, as none of the transformants tested exhibit any phage release (Figure 4-8B). Similarly, spontaneous phage release from an EagleEye lysogen is substantially reduced after pMO01 selection, suggesting some degree of incompatibility with EagleEye *parABS*. In contrast, there is no apparent effect of pMO01 selection on L5 phage release, indicating pMO01 exhibits no incompatibility with the integrating phage. When the same experiment is performed using pLO87, which lacks *parABS*, there is no apparent reduction in RedRock or EagleEye phage release, indicating the *parABS* locus on pMO01 is responsible for the incompatibility (Figure 4-8B).

4.3.8 RedRock and Alma ParB homologs exhibit distinct specificities

The degree of diversification among Cluster A partitioning systems is not obvious. Cluster A ParA proteins are grouped in pham 7133 and are monophyletic (Figure 4-3A). However, the ParB proteins are more diverse, grouped into five phams, and exhibit more distinct phylogenetic clades (Figure 4-3B). Additionally, the survey of *parS* loci suggests there may be some degree of diversification of ParB binding specificity (Figure 4-2). For example, many phages in Subclusters A6 and A11 carry ParB proteins with substantial similarity to RedRock ParB. In contrast, Echild and 40AC ParB proteins are distantly related to the other Cluster A homologs.

We selected several partitioning systems that represent the spectrum of ParB diversity, and Graham Hatfull performed a detailed analysis of their cognate *parS* loci (Figure 4-9A). Similar to the RedRock system, the partitioning systems in *Mycobacterium* phages Gladiator (Subcluster A6), Alma (Subcluster A9), and Echild (Subcluster A2), as well as in *Gordonia* phage KatherineG (Subcluster A15), contain tandem repeats of 6-8 bp sequences in their *parS-L*

and *parS-R*. The consensus *parS* sequences correlate with ParB diversity: RedRock *parS* is more similar to Gladiator and Alma *parS* than to Echild and KatherineG *parS* (Figure 4-9A).

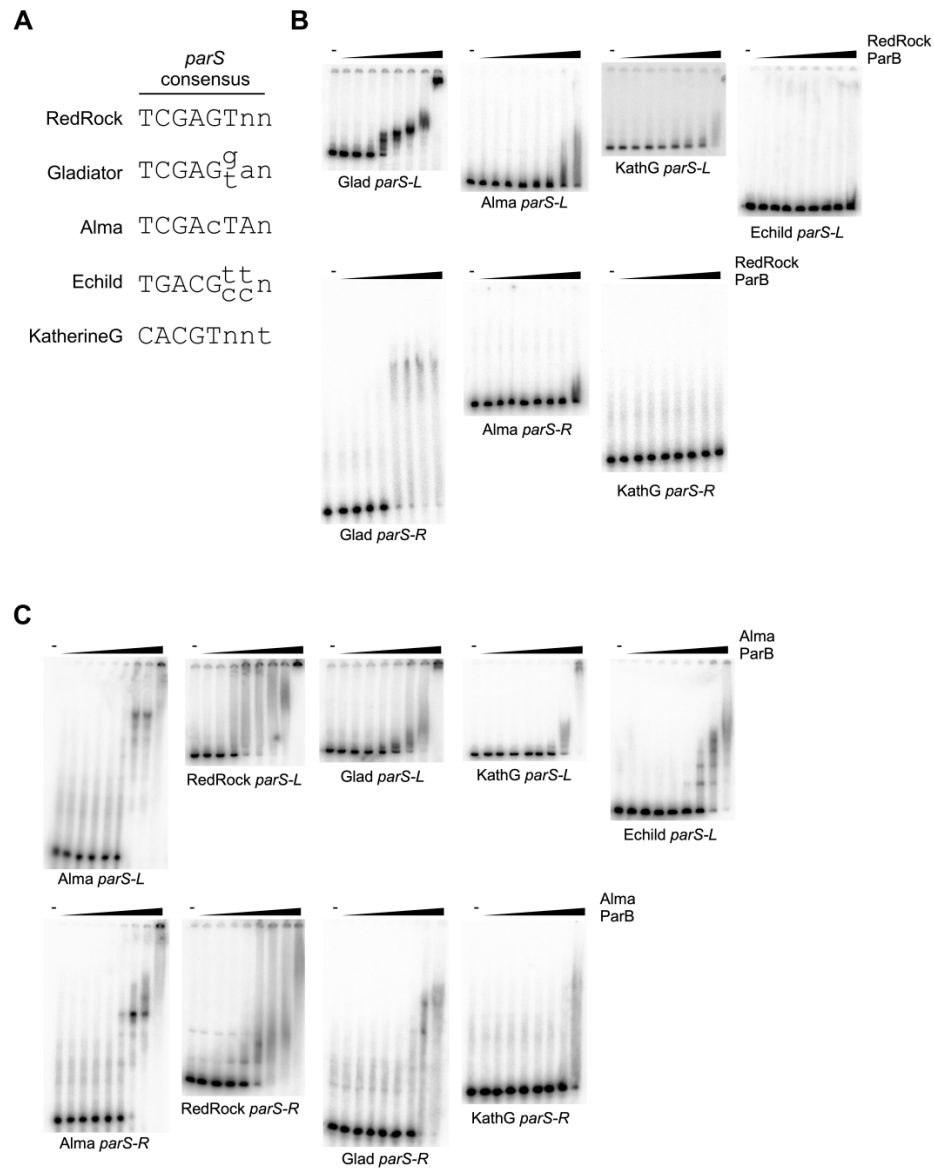


Figure 4-9. RedRock and Alma ParB exhibit distinct *in vitro* *parS* binding affinities.

(A) Manual alignment by Graham Hatfull of consensus sequences representing manually identified *parS* sites in several Cluster A prophages: Gladiator (Glad), Alma, KatherineG (KathG), RedRock, and Echild. The binding affinities of purified (B) RedRock ParB and (C) Alma ParB were compared for different substrates. Serially diluted ParB was incubated with radiolabeled *parS-L* and *parS-R* substrates. This experiment was performed by Wei Ng, and Rebekah Dedrick. Figure is adapted from the original generated by Graham Hatfull (Dedrick et al., 2016).

Juan Cervantes Reyes, Wei Ng, and Rebekah Dedrick tested whether RedRock ParB exhibits distinct specificities for cognate versus non-cognate *parS* sites (Figure 4-9B). RedRock ParB is able to bind strongly to a substrate containing Gladiator *parS-L*, and exhibits some binding of a substrate containing Gladiator *parS-R*. In contrast, it is unable to recognize the *parS-L* and *parS-R* of Alma or KatherineG, or the *parS-L* of Echid. Alma ParB is more distantly related to RedRock ParB, suggesting it may exhibit unique binding specificity (Figure 4-3B). Wei Ng, and Rebekah Dedrick purified Alma ParB and tested its binding affinity for various *parS* substrates (Figure 4-9C). It exhibits tight binding to Alma *parS-L* and *parS-R*, and similar to RedRock ParB, it does not exhibit much affinity for KatherineG *parS* substrates. However, in contrast to RedRock ParB, it exhibits some binding to Echid and Gladiator *parS* substrates. It also exhibits some binding to RedRock *parS* substrates, even though RedRock ParB does not exhibit any binding to Alma *parS* substrates (Figure 4-9B, C). These results suggest that RedRock and Alma partitioning systems, although not identical, do genetically interact.

4.3.9 *parABS* systems confer prophage incompatibility

Since Alma ParB recognizes RedRock *parS* *in vitro*, these phages may exhibit *parABS*-mediated incompatibility *in vivo*, similar to plasmids. Since they are heteroimmune, I tested whether they exhibit incompatibility using a double lysogeny assay (Figure 4-10A). A RedRock lysogen was superinfected with either Alma, Pioneer (also in Subcluster A9 as Alma but contains a more distant *parABS* system), and Bxb1 (a heteroimmune Subcluster A1 integrating phage). Colonies of *M. smegmatis* that grew from the superinfected clearings were picked, purified and grown in liquid culture. During this process, isolates were tested for whether they were a single

lysogen of the defending RedRock prophage, a single lysogen of the superinfecting phage, or a double lysogen by patching colonies or spotting cultures on lawns of each single lysogen.

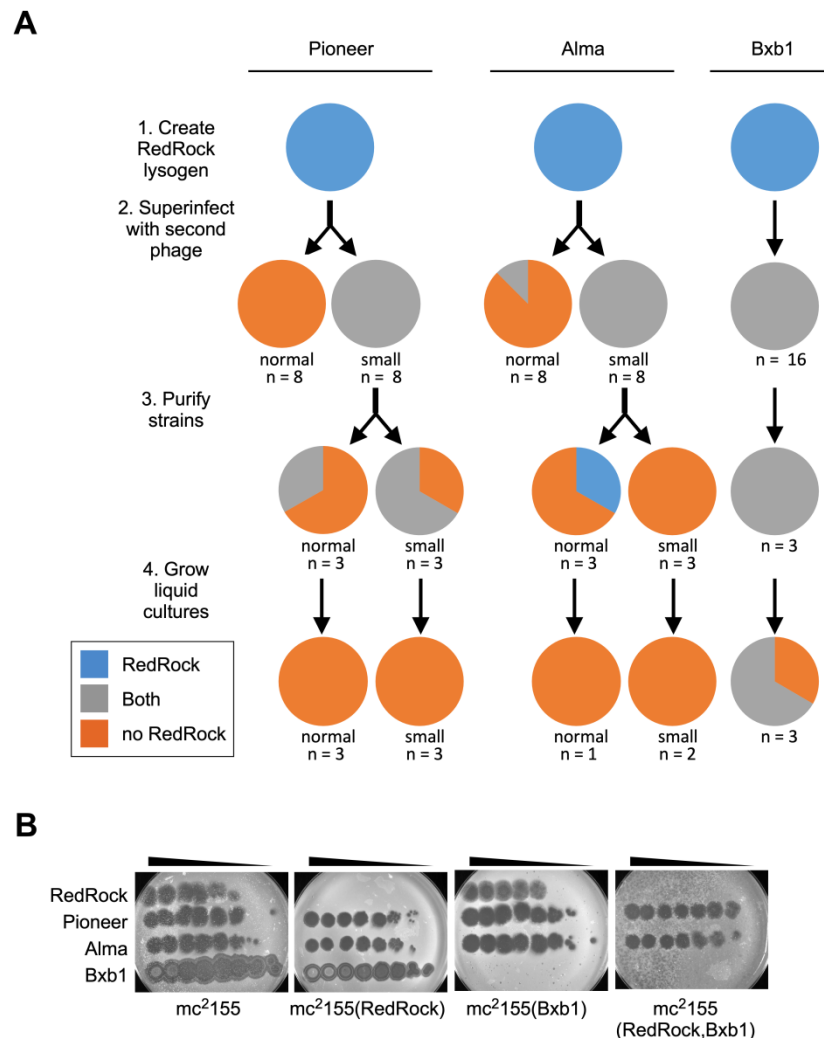


Figure 4-10. Partitioning systems promote prophage-prophage incompatibility.

(A) Experimental design to test for partitioning-mediated incompatibility between two extrachromosomal prophages. (Step 1) A RedRock lysogen was generated and (Step 2) superinfected with Alma, Pioneer, or Bxb1. (Step 3) Potential double lysogens were clonally purified and (Step 4) cultured in liquid medium to saturation. At each stage strains were tested for double lysogeny. Sample sizes (n) and colony morphologies (“normal” or “small”) at each stage are indicated, and pie charts reflect fractions of each type of single or double lysogen identified. **(B)** Superinfection immunity assays of several phages against mc²155, single lysogens (RedRock or Bxb1), or a double lysogen (RedRock and Bxb1). Black triangles indicate 10-fold serial dilutions of phage lysate. Figure adapted from (Dedrick et al., 2016).

After superinfection with Bxb1, all colonies during purification and most liquid cultures display spontaneous phage release on RedRock and Bxb1 lysogens, indicating they are double lysogens (Figure 4-10A). Double lysogeny is also confirmed by superinfection immunity assay (Figure 4-10B). Similar to the corresponding single lysogens, the double lysogen is not immune to Pioneer or Alma. But unlike the single lysogens, the double lysogen is immune to both RedRock and Bxb1. Unlike Bxb1, superinfection with Pioneer or Alma produces different results. Double lysogens can be detected at earlier stages of purification, but eventually all isolates are single lysogens. The majority of these single lysogens carry Pioneer or Alma prophages (Figure 4-10A). These results suggest that the partitioning systems in Pioneer and Alma exhibit incompatibility with the RedRock partitioning system, and superinfection of a RedRock lysogen results in the displacement of RedRock.

4.3.10 Evolution of Cluster A *parABS* systems

CBP proteins exhibit greater diversity than NTPase proteins, which may enable replicons to avoid incompatibility in the presence of competing systems (Fothergill et al., 2005; Gerdes et al., 2000). NTPases may not be subject to the same selective forces since they indirectly interact with the replicon through interactions with the CBP. The phylogenies of Cluster A partitioning genes suggest that ParA and ParB evolve differently. In general, ParA proteins are more closely related to each other, exhibiting smaller branch lengths than ParB proteins (Figure 4-3). If these two genes evolve differently, they may exhibit different rates of evolution, reflected by the number of synonymous (K_S) and nonsynonymous (K_A) mutations present between two homologs. K_A/K_S ratios less than one indicate genes are subject to purifying selection. Ratios

greater than one indicate genes are subject to diversifying selection. Ratios that approximate one indicate neutral selection (Kimura, 1968; Swanson and Vacquier, 1998).

To explore the evolutionary patterns of Cluster A partitioning systems, K_A/K_S ratios were measured for all pairwise comparisons of ParA and ParB homologs among 27 *parABS* loci (Figure 4-11, see Materials and Methods). In general, ParA is subject to strong purifying selection, with most K_A/K_S ratios < 0.3 . In contrast, there appears to be relaxed selection on ParB, with K_A/K_S ratios exhibiting a wider range that approaches 1. Although ParB does not exhibit evidence of diversifying evolution, it does appear to be evolving at a faster rate than its ParA counterpart, consistent with predictions, suggesting it diversifies to avoid genetic conflicts with other systems.

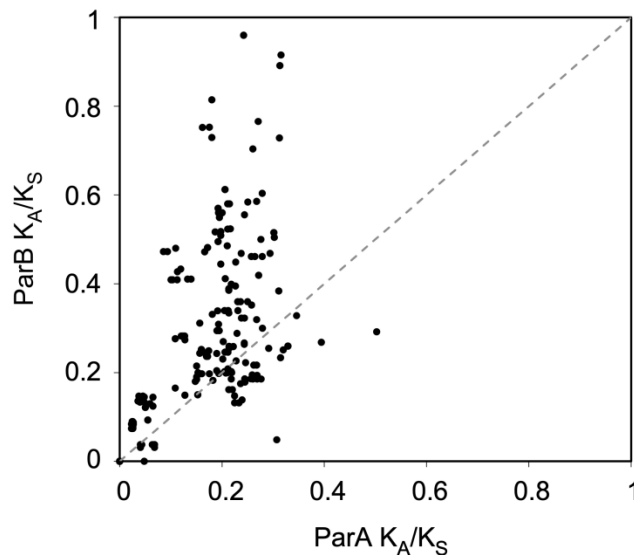


Figure 4-11. ParB evolves at a different rate than ParA.

Scatter plot comparing all pairwise K_A/K_S ratios of cognate *parA* and *parB* unique gene sequences from 27 actinobacteriophage partitioning systems (Supplementary Table 4-1). The $y = x$ line is plotted for reference. Figure adapted from (Dedrick et al., 2016).

4.3.11 Investigation of RedRock prophage origin of replication

Phages that integrate into the host genome are passively replicated along with the rest of the host genome by the host's replication machinery. In contrast, extrachromosomal replicons, such as plasmids and prophages, often replicate independent of the host, but this requires an origin of replication as well as specific replication genes (del Solar et al., 1998; Weigel and Seitz, 2006). Characterization of these replication strategies can enable the development of new vectors. Skews in DNA composition have been used to identify bacterial origins of replication (Lobry, 1996; Worning et al., 2006) as well as to analyze viral genomes, such as enterobacteria phage SP6 (Chew et al., 2007; Da Silva and Upton, 2005; Dobbins et al., 2004; Grigoriev, 1998, 1999). Many factors can impact skews in nucleotide composition, including replication, transcription, and translation (Grigoriev, 1999; McLean et al., 1998; Thomas et al., 2007; Tillier and Collins, 2000). Diverse skew metrics, such as third codon position skews (McLean et al., 1998; Mrazek and Karlin, 1998) or the Z curve (Zhang et al., 2003), have been used to identify different types of genomic elements (Grigoriev, 1998). Sometimes, skews can help to evaluate prophage origins of replication, such as the correlation of GC skews with the origin of replication for prophage P1 (Figure 4-12A). However, skews may not produce as robust a signature for the origin of replication for other prophages, such as N15 (Figure 4-12A).

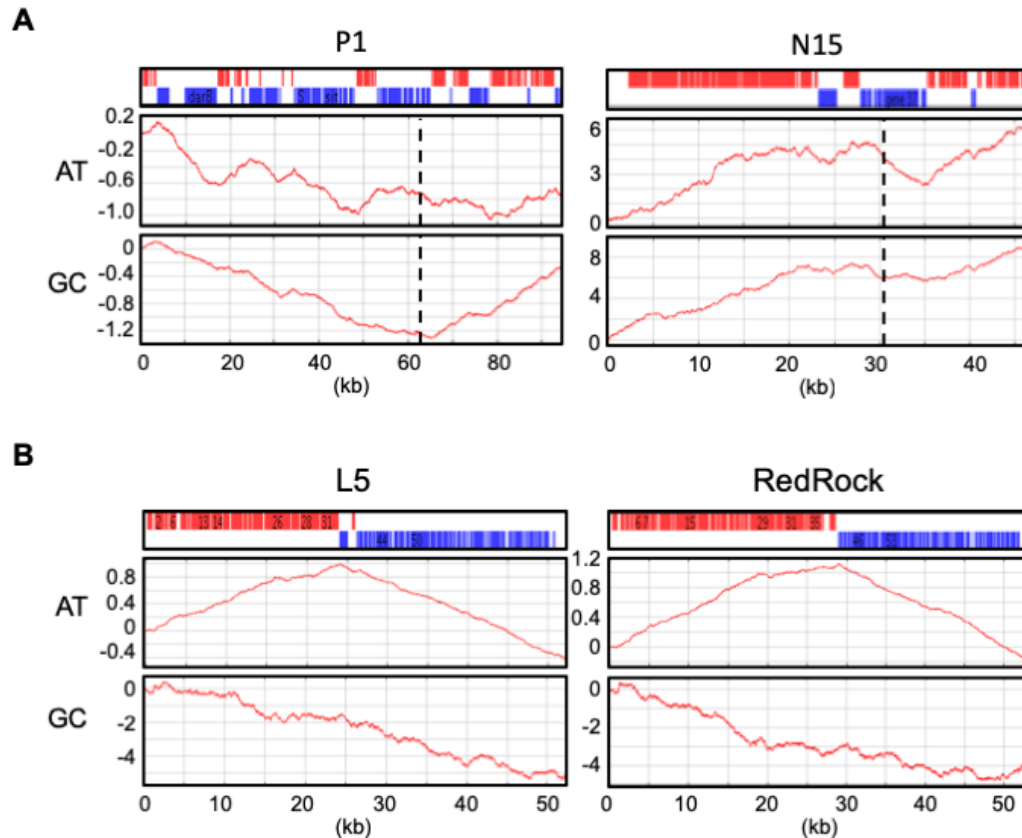


Figure 4-12. Cluster A inheritance strategy does not impact nucleotide composition.

Cumulative AT and GC DNA skews were computed for the genomes of **(A)** extrachromosomal enterobacteria phages (P1 and N15) and **(B)** integrating (L5) and extrachromosomal (RedRock) Subcluster A2 *Mycobacterium* phages. Genome maps display rightward and leftward transcribed genes as red and blue boxes, respectively. Dashed lines mark the genomic location of empirically determined prophage origins of replication. Figure adapted from (Dedrick et al., 2016).

Little is known about mechanisms of replication for *Mycobacterium* phages during lytic growth or lysogeny. AT composition for Cluster A integrating phages, such as L5, are very skewed (Figure 4-12B). However, the skew coincides with direction of transcription on each arm of the genome, suggesting that they are the result of transcriptional or translational processes instead of replication processes. Prophage inheritance strategy may impact DNA composition, and differences in DNA skews between extrachromosomal and integrating Cluster A phages may highlight possible origins of replication during lysogeny. However, comparison of various

compositional skews in RedRock, such as AT and GC skews, are not substantially different from skews in L5 (Figure 4-12B). Therefore, DNA skews are unable to help evaluate Cluster A extrachromosomal phage replication.

Although DNA composition does not provide insight into Cluster A prophage replication strategies, expression patterns may provide insight. Plasmid and prophage replication systems are diverse, and some are dependent on transcription across the origin of replication to initiate replication (Weigel and Seitz, 2006). Interestingly, during lysogeny and early lytic growth, expression is observed in RedRock upstream of *parA* on the bottom strand across a locus with no predicted gene (Figure 4-5B). Similar expression patterns are observed at the syntenic position in Alma, Pioneer, and EagleEye (Figure 4-5C). Matt Olm evaluated this intergenic region in RedRock with plasmid constructs and his results suggest there is a promoter opposite of *parA* that, in contrast to P_{par} , is up-regulated by ParB (Dedrick et al., 2016). These data suggest that the Cluster A *parABS* cassette contains an additional promoter, P_{ori} , oriented opposite of P_{par} , and it may be involved in transcription-dependent replication (Figure 4-1B).

To test whether the *parABS* locus contains an origin of replication, Rachael Rush and I created the plasmid pRR06 by cloning the RedRock *parABS* cassette into pMD02, a vector that does not contain a mycobacterial origin of replication or integration cassette and is thus not competent for replication in *M. smegmatis* (Bibb and Hatfull, 2002). However, when pRR06 was transformed into *M. smegmatis*, no transformants were recovered (see Materials and Methods). These results suggest that there may not be a prophage origin of replication at the *parABS* locus.

4.3.12 Analysis of other partitioning systems

Phages Echild (Subcluster A2) and 40AC (Subcluster A17) contain partitioning cassettes that are more distantly related to the cassettes of the other 39 Cluster A phages (Figure 4-3). As with the other Cluster A ParA homologs, their ParA proteins contain Type Ib NTPase-related domains, and they are most closely related to ParA proteins in *Gordonia* phages Soups, Rosalind, and KatherineG (Figure 4-3A). However, relative to other Cluster A ParB proteins, Echild ParB contains few Type Ib CBP-related domains, and 40AC ParB does not contain any CBP domains at all (Figure 4-3B). Additionally, 40AC does not contain as easily identifiable *parS* sites (Figure 4-2). Although an Echild lysogen could be generated by clonal purification on solid Middlebrook 7H10 medium, the prophage does not appear to be stably maintained in liquid culture using Middlebrook 7H9 medium. These results suggest that the partitioning systems in Echild and 40AC function differently than other Cluster A partitioning systems or that they are not fully functional in *M. smegmatis*.

Using the phylogenetic analysis of NTPase and CBP proteins, the type of partitioning systems that have been reported in other types of phages can be investigated. The partitioning system of *Yersinia* phage PY54 (Hertwig et al., 2003) is likely to be categorized as Type Ia, since its NTPase (SpyA) and CBP (SpyB) are closely related to the NTPase (SopA) and CBP (SopB) in enterobacteria phage N15 (Ravin, 2015). Similarly, *Halomonas* phage ϕ HAP-1 (Mobberley et al., 2008) and *Vibrio* phage Vp58.5 (Zabala et al., 2009) contain partitioning systems that are likely categorized as Type Ib. Their predicted NTPases contain Type Ib domains, and although they contain an adjacent gene with no predicted function, both gene products contain domains related to Type Ib CBPs (Figure 4-3).

Several partitioning systems may function differently than previously characterized systems. The *Streptomyces* phage pZL12 (Zhong et al., 2010) is a singleton and is the only other actinobacteriophage with a predicted partitioning system other than phages grouped in Cluster A, but it is not well characterized (Supplementary Table 4-1). The NTPase and CBP gene products in its partitioning system are related to the NTPase in the *Streptomyces* plasmid pSLE1 (Gomez-Escribano et al., 2015). The NTPases in both systems contain many Type Ib NTPase-related domains (Figure 4-3A). However, the CBPs in both systems contain many Type Ia CBP-related domains (Figure 4-3B). Overall, pZL12 may have acquired its partitioning system from a *Streptomyces* plasmid, and it cannot clearly be grouped into Type Ia or Ib.

Several extrachromosomal replicons identified in *Leptospira* hosts, lcp1, lcp2, and lcp3 have been predicted to have partitioning cassettes, but they have not been well characterized (Zhu et al., 2015). Only lcp3 is reported as a prophage, and the closest relative to its ParA is the ParA in *Leptospira* phage LE1 (Bourhy et al., 2005), both of which have Type Ib domains (Figure 4-3A). The ParB for lcp3 and LE1 do not have any CBP domains, but LE1 ParB has been shown to play a role in plasmid stability (Bourhy et al., 2005). Therefore, lcp3 and LE1 may utilize partitioning systems with novel genetic components.

The NTPase of *Vibrio* phage pVv01 (Hammerl et al., 2014) contains Type Ib domains, but the adjacent gene has no prediction function, and its gene product does not contain any domains related to previously characterized CBPs (Figure 4-3). This may not be a functional partitioning system, or it may function much differently than other characterized systems.

4.4 DISCUSSION

The majority of temperate actinobacteriophages utilize integration systems, so the evolutionary benefits of partitioning systems are not clear. Since genetically related phages grouped in Cluster A can utilize either system, the prophage inheritance strategy may not be dependent on strategies for infection, lytic growth, or maintenance of lysogeny with the superinfection immunity system. Integration systems are only valuable if the host harbors a compatible integration site, so there may be selective pressure to remain extrachromosomal within particular hosts. Although the Cluster A partitioning systems eliminate the dependence on host integration sites, they may instead exhibit a greater dependence on host genes required for extrachromosomal prophage replication. Inheritance strategy may also be useful to enhance prophage stability in the presence of other prophages. When multiple prophages integrate into the same host integration site, they may be destabilized (Bertani, 1971). Acquisition of partitioning systems may alleviate the burden imposed on particular integration sites. However, since extrachromosomal prophages utilizing similar partitioning systems are also destabilized, residing as an extrachromosomal state does not completely avoid the problem. Additionally, some extrachromosomal prophages, such as Echild, ArcherNM, and LadyBird appear to be unstable even in the absence of a second prophage, suggesting that partitioning systems do not guarantee efficient prophage inheritance.

New genetic tools can be developed from these diverse systems to enhance mycobacterial genetics. Few mycobacterial vectors have been well characterized (Bachrach et al., 2000; Gavigan et al., 1997; Labidi et al., 1985), and the 42 extrachromosomal actinobacteriophages represent a collection of diverse partitioning systems that could be exploited to develop new vectors. All Cluster A partitioning systems are likely derived from a common evolutionary

origin, and RedRock, Alma, and Pioneer harbor very similar systems. However, several phages may carry more distantly related, orthogonal systems. The partitioning systems in Cluster A *Gordonia* phages (Subcluster A15) appear distinct, but it is not obvious whether these function in mycobacteria. Some of the mycobacterial partitioning systems that are most distantly related to RedRock and Alma are present in Echid and 40AC (Subcluster A2), Luchador (Subcluster A14), and Loser and CRB1 (Subcluster A2). Additionally, the partitioning system in *Streptomyces* phage pZL12 is completely different than Cluster A systems. If the Cluster A *parABS* systems contain an origin of replication, they could enable the development of orthogonal mycobacterial plasmid constructs, but further work needs to be done to characterize how Cluster A prophages replicate during lysogeny. Nevertheless, the Cluster A *parABS* systems can be used to improve mycobacterial plasmid stability.

5.0 FUNCTION AND EVOLUTION OF CLUSTER A IMMUNITY SYSTEMS

The majority of data in this chapter regarding evolution of superinfection immunity are unpublished, and I performed all experiments and analyses, with the exception of sequencing RNA samples, sequencing mutant phage DNA, and assembling mutant phage genomes, which were performed by Dan Russell and Rebecca Garlena. The data in this chapter pertaining to evolution of D29 and its relatives was published in *BMC Microbiology* (Dedrick et al., 2017b). For this latter project, RNA isolation and sequencing were performed by Rebekah Dedrick, Dan Russell, and Rebecca Garlena. I performed the transcriptome analysis and all other experiments.

5.1 INTRODUCTION

Although phages must evolve to overcome defenses imposed by their bacterial hosts, they must also evolve to overcome other phages that are competing for the same resources (Dedrick et al., 2017a; Doron et al., 2018). Many phages are temperate (Hatfull, 2010), but the selective forces for lysogeny are not well understood (Chibani-Chennoufi et al., 2004), and although lysogeny may be evolutionarily beneficial, the host remains susceptible to a second round of infection by other genetically identical (homotypic), similar (mesotypic), or unrelated (heterotypic) phages (Berngruber et al., 2010; Chibani-Chennoufi et al., 2004; Refardt, 2011). As a result, temperate phages must evolve mechanisms to control lysogeny while also defending against other superinfecting phages and escaping other prophage defenses.

The genetic circuits prophages use to maintain lysogeny, their immunity systems, are a target of these evolutionary forces. The evolutionary process in which homotypic immunity systems diverge into distinct heterotypic specificities is also poorly understood. Superinfection homoimmunity and heteroimmunity are typically symmetric binary phenotypes, in which reciprocal prophage-phage interactions produce the same phenotype of either complete defense, or complete absence of defense. In the λ system, co-evolution of CI specificity and the operators may drive the evolution of immunity groups (Berngruber et al., 2010; Campbell, 1994). However, since multiple regulatory elements must co-evolve to ensure the regulatory circuit remains functional, the transition to a new immune specificity is unlikely to occur in a single step (Campbell, 1994). Within a group of genetically related phages, it is not clear how intermediate, mesotypic specificities impact the symmetric binary phenotype. Additionally, we do not know the degree to which mutations conferring homotypic virulence also confer mesotypic virulence.

Mycobacterium phage L5 and its genetic relatives represent a well-defined collection of phages that exhibit sufficient genetic diversity to examine how immunity systems evolve. Here, I sought to investigate how Cluster A superinfection immunity and virulence are impacted by immunity system evolution. The evolutionary divergence of Cluster A immunity systems results in a group of mesoimmune phages that exhibit a complex network of repressor-mediated asymmetric and incomplete immunity phenotypes. The evolutionary transition from homoimmunity to heteroimmunity may be nonlinear and frequent. Defense escape mutants show that immunity systems can be shaped by both homoimmune and mesoimmune phages, and mutations confer varying degrees of homotypic and mesotypic virulence. Additionally, I show that expression from the P_{left} locus and expression of several genes downstream from P_{rep} are toxic to the host.

5.2 MATERIALS AND METHODS

5.2.1 Phamerator database construction

For the majority of bioinformatic analyses in this chapter, the database *Actinobacteriophage_1321* was created using Phamerator (Cresawn et al., 2011), consisting of 1,305 manually annotated genomes of actinobacteriophages isolated from the environment and 16 engineered or isolated mutants (described below). To investigate the evolutionary dynamics of D29 and its relatives, the database *Actinobacteriophage_937* was created, consisting of 937 manually annotated actinobacteriophages. Genes are grouped based on amino acid sequence similarity using kClust (Hauser et al., 2013) implemented in the Phamerator pipeline (Appendix A). The databases are available online (http://phamerator.webfactional.com/databases_Hatfull).

5.2.2 Identification and analysis of stoperator sequences

Stoperator sequences were automatically identified in nearly all Cluster A genomes using MEME (Bailey et al., 2009) using the following parameters: site distribution = any number of repetitions, maximum motifs = 2, motif length = 12-16 bp, number of sites = 50, search both strands = yes. For each genome, the motif that most closely resembled empirically determined L5 and Bxb1 stoperator sites was selected. All motifs were converted to the sense strand and manually aligned. Motif logos representing the aligned sequences were created with Weblogo (Crooks et al., 2004). Stoperator sequences were compared in R using *Biostrings* and *TFBSTools* packages (Tan and Lenhard, 2016). Position weight matrices (PWMs) of the core 13 bp sequence were created using the *PFMatrix* and *toPWM* functions, using the *log2probratio* method and

default values for background and pseudocount settings. Pairwise PWM normalized Euclidean distances were computed using the *PWMSimilarity* function and larger distances represent more dissimilar PWMs (Harbison et al., 2004; Tan and Lenhard, 2016). Stoperators were determined to be oriented in the direction of transcription (syn-oriented) if they were located on the top strand to the left of the genome center or on the bottom strand to the right of the genome center. The center of the genome was defined as the coordinates of the *integrase* (for integrating phages) or *parA* (for extrachromosomal phages) gene. To generate genomic distributions of stoperators in L5 clade phages, coordinates of all stoperators in each phage were adjusted relative to the coordinates of the genomic feature of interest in that specific phage, and histograms were created using adjusted coordinates for all L5 clade phages.

5.2.3 Computation of genomic similarity metrics

Pairwise whole genome nucleotide distance between all phage genomes were computed using Mash, and pairwise gene content dissimilarity (referred to as gene content distance in this chapter) were computed using kClust-based phams, as described in Chapter 2. For pairs of phages, gene content dissimilarity ranges from 0 (all gene phams are identical) to 1 (no gene phams are identical), and nucleotide distance ranges from 0 (identical sequence) to 0.5 (unrelated sequence).

5.2.4 Genetic distance of specific Cluster A genes

Amino acid sequences for 336 full-length homologs of the Cluster A immunity repressor present in the database (grouped into pham 3247, 38916, or 38877) were aligned using MAFFT (Kato et al., 2002). The N-terminus of the alignment was manually trimmed in SeaView (Gouy et al., 2010), and the trimmed alignment was split into N-terminal and C-terminal domains as previously reported (Ganguly et al., 2007). Uncorrected distances between taxa in the full length, N-terminus, and C-terminus alignments were computed using the EMBOSS *distmat* tool with no gap weight and reported as the number of substitutions per 100 amino acids (<https://www.ebi.ac.uk/Tools/emboss/>). The 20 amino acid helix-turn-helix domain was identified in all taxa from the MAFFT alignment based on previous reports (Jain and Hatfull, 2000; Pope et al., 2011b). Uncorrected distances for full-length proteins of 311 homologs of Cas4 (pham 29663), 306 homologs of EndoVII (pham 39443), 311 homologs of DNA Polymerase (pham 39585), and 311 homologs of Portal (pham 38438) genes present in Cluster A phages were computed in the same way. Uncorrected hamming distances between helix-turn-helix domains were computed using the *stringdist* R package. Unlike whole genome distances and stoperator motif distances, gene-specific distances are limited to phages that carry a homolog of the gene of interest.

5.2.5 Repressor nucleotide alignment and phylogeny

Nucleotide sequences for 79 immunity repressors present in phages from the L5 clade were aligned with webPRANK (Loytynoja and Goldman, 2010), a phylogenetic tree was constructed using maximum likelihood in SeaView (Gouy et al., 2010), and it was annotated using Evolview (Zhang et al., 2012a).

5.2.6 Preparation of phage lysates and lysogens

A diverse set of phages was selected for immunity assays, representing multiple subclusters, utilizing different prophage inheritance strategies (integration or extrachromosomal), and carrying complete or mutant repressor genes (Table 5-1). All phages used for immunity assays were plaque purified at least twice and confirmed by PCR using primers that amplify near the right end of the genome, in which there is substantial sequence diversity among Cluster A phages. Primer pairs used for each phage are as follows: Alma (oTM156, oTM157), ArcherNM (oTM144, oTM145), Bxb1 (oTM114, oTM115), phiTM33 (oTM72, oTM73), D29 (oTM179, oTM180), DarthPhader (oTM237, oTM238), DaVinci (oTM74, oTM75), Drake55 (oTM253, oTM254), Dreamboat (oTM152, oTM153), EagleEye (oTM146, oTM147), Echid (oTM158, oTM159), Et2Brutus (oTM76, oTM77), Gladiator (oTM78, oTM79), Jaan (oTM255, oTM256), Jeffabunny (oTM80, oTM81), Journey13 (oTM245, oTM246), L5 and L5 derivatives (oTM69, oTM70), LadyBird (oTM120, oTM121), Larenn (oTM251, oTM252), MissWhite (oTM231, oTM232), Mulciber (oTM82, oTM83), Peaches (oTM154, oTM155), Petruchio (oTM138, oTM139), Pioneer (oTM160, oTM161), Piro94 (oTM124, oTM125), RedRock (oTM126, oTM127), Serenity (oTM243, oTM244), StarStuff (oTM162, oTM163), Trixie (oTM86,

oTM87), Updawg (oTM247, oTM248), phiTM1 and phiTM4 HA-tagged *rep* (oTM55, oTM60), and phiTM6 FLAG-tagged *rep* (oTM55, oTM59)(Appendix B). Lysates were expanded one round from a plaque pick by plating phage with mc²155, incubating at 37°C for 24-36 h, incubated with 5 ml Phage Buffer at room temperature for 4-5 h, and filtered.

Lysogens were created by spotting high titer phage lysates on a lawn of mc²155, picking cells from the center of the spot after 3-7 days, and picking colonies, which were subsequently clonally purified at least two times. Strains were confirmed as lysogens by PCR using the same primers as for lysate confirmation, by verifying that cells exhibit spontaneous phage release when spotted onto a lawn of mc²155, and by verifying that the strain is immune to infection from the parent phage. Lysogens for phages Echid (Chapter 4), Journey13, and Piro94 could not be generated (Table 5-1). Lysogenization of some phages was not tested (Table 5-1).

Table 5-1. Phages used in this chapter.

Phage Name	Sub-cluster	Inheritance	Lysogen generated	Parent phage	Mutant origination	Repressor mutation	Other mutations
phiTM45	A1	<i>Int</i>	<i>N/A</i>	Bxb1	Bxb1 CRS DEM	Nonsense mutation	<i>N/A</i>
phiTM33	A2	<i>Int</i>	no	Che12	unintentional isolate	N-terminus deletion	Gene 29 F223L
phiTM46	A6	<i>parABS</i>	<i>N/A</i>	DaVinci	Gladiator CRS DEM	1 bp insertion	<i>N/A</i>
phiTM36	A16	<i>parABS</i>	<i>N/A</i>	EagleEye	Pioneer lysogen DEM	Complete deletion	2.7 kb deletion in right arm
phiTM39	A11	<i>parABS</i>	<i>N/A</i>	Et2Brutus	L5 lysogen DEM	Nonsense mutation	Holin V9G; Gene 98 sense mutation
phiTM40	A11	<i>parABS</i>	<i>N/A</i>	Et2Brutus	Trixie lysogen DEM	A38V in HTH domain	Holin I15S
phiTM47	A6	<i>parABS</i>	<i>N/A</i>	Gladiator	Gladiator CRS DEM	1 bp insertion	<i>N/A</i>
phiTM41	A2	<i>Int</i>	yes	L5	Trixie lysogen DEM	<i>N/A</i>	Gene 89 F47L
phiTM35	A9	<i>parABS</i>	<i>N/A</i>	Pioneer	EagleEye lysogen DEM	Nonsense mutation	<i>N/A</i>
phiTM42	A2	<i>Int</i>	<i>N/A</i>	Trixie/RedRock	Trixie lysogen DEM	1 bp insertion	Hybrid of Trixie/RedRock genomes
phiTM43	A2	<i>Int</i>	No	D29	unintentional isolate	same as D29	Gene 32 P202S
phiTM44	A2	<i>Int</i>	No	D29	unintentional isolate	same as D29	Gene 32 P202T; Gene 59.2 sense mutation
phiTM38	A2	<i>Int</i>	<i>N/A</i>	phiTM44	Et2Brutus lysogen DEM	Point mutation	<i>N/A</i>
phiTM1	A2	<i>Int</i>	yes	L5	BRED	C-terminal HA tag	<i>N/A</i>
phiTM4	A2	<i>Int</i>	no	phiTM1	unintentional isolate	C-terminal HA tag	Gene 70 A145E
phiTM6	A2	<i>Int</i>	yes	L5	BRED	C-terminal FLAG tag	<i>N/A</i>
Jeffabunny	A6	<i>parABS</i>	no	<i>N/A</i>	<i>N/A</i>	Complete deletion	<i>N/A</i>
MissWhite	A2	<i>Int</i>	no	<i>N/A</i>	<i>N/A</i>	Complete deletion	<i>N/A</i>
D29	A2	<i>Int</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	N-terminus deletion	<i>N/A</i>
Bxb1	A1	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Dreamboat	A1	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Petruchio	A1	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Et2Brutus	A11	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Mulciber	A11	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
DarthPhader	A12	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
EagleEye	A16	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Echild	A2	<i>parABS</i>	no	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Journey13	A2	<i>Int</i>	no	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Piro94	A2	<i>Int</i>	no	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
ArcherNM	A2	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Drake55	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

Jaan	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
L5	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
LadyBird	A2	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Larenn	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
RedRock	A2	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Serenity	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
StarStuff	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Trixie	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Updawg	A2	<i>Int</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Peaches	A4	<i>Int</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
DaVinci	A6	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Gladiator	A6	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Alma	A9	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
Pioneer	A9	<i>parABS</i>	yes	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>

5.2.7 RNAseq

Strand-specific transcription profiles of Et2Brutus, Gladiator, and Trixie lysogens were measured as described in Chapter 4 and viewed using Integrative Genomics Viewer (IGV)(Thorvaldsdottir et al., 2013). I isolated RNA and prepared samples, Dan Russell and Rebecca Garlena sequenced the samples, and I analyzed the data. Raw fastq data are deposited in the Gene Expression Omnibus (GEO) under accession GSE123612.

For strand-specific transcription profiling of phages L5, StarStuff, and D29 during lysogeny and infection, Rebekah Dedrick isolated RNA and prepared samples, Dan Russell and Rebecca Garlena sequenced the samples, and I analyzed the data, as described in Chapter 4. Raw fastq data are deposited in the Gene Expression Omnibus (GEO) under accession GSE99182.

5.2.8 Repressor overexpression and EMSAs

Rep_{Trixie} was amplified from the Trixie genome (coordinates 45,599-46,174) with primers oTM13 and oTM14, cloned into the expression vector pET-21a using the *Nde*I and *Hind*III sites to create the plasmid pTM1, which carries Rep_{Trixie} C-terminally tagged with His and a short linker (KLAAALEHHHHHH). pTM1 was transformed into NEB5 α cells. Sequence-verified plasmid constructs were transformed into BL21 STAR(DE3) cells and single colonies were grown in LB medium supplemented with carbenicillin. Repressor expression was induced with 1 mM IPTG for 3 h, cells were lysed by resuspending in Lysis Buffer (50 mM Tris pH 8.0, 300 mM NaCl, 10% glycerol), treating with 1 mg/mL lysozyme for 30 min on ice, and light sonication (Villanueva et al., 2015). C-terminally His-tagged Rep_{Trixie} was purified using a

Nickel-NTA matrix and dialyzed overnight with Storage Buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.1 mM EDTA, 0.1 mM DTT, 50% glycerol). DNA substrates for electrophoretic mobility shift assays (EMSAs) were designed to be 30 bp long, consisting of a 13 bp stoperator sequence flanked by 8-9 bp of sequence. Complementary 30 bp oligonucleotides were synthesized, radiolabeled at the 5' end with γ - ^{32}P and annealed (Villanueva et al., 2015). Oligos for each substrate are as follows: Alma (oTM21, oTM22), Gladiator (oTM23, oTM24), Peaches (oTM17, oTM18), RedRock (oTM31, oTM32), Rockstar (oTM19, oTM20), Trixie (oTM33, oTM34), C₉ (oTM43, oTM44), C₉G₁₀ (oTM41, oTM42), C₉C₁₁ (oTM39, oTM40), C₉A₁₂ (oTM37, oTM38), C₉G₁₀C₁₁ (oTM49, oTM50), C₉G₁₀A₁₂ (oTM47, oTM48), C₉C₁₁A₁₂ (oTM45, oTM46), C₉G₁₀C₁₁A₁₂ (oTM35, oTM36), L5 gene *3I* negative control (oTM29, oTM30). The sequences of the 30 bp substrates reflecting homologous stoperator sites and the L5 gene *3I* negative control completely match the genome sequence. For the 30 bp substrates in which the Trixie stoperator is progressively converted to a Peaches stoperator site, the variable 13 bp sequence is flanked by invariable 8-9 bp derived from the Trixie substrate. EMSAs were performed with serially diluted Rep, electrophoresed on an 8% polyacrylamide gel, and imaged, as previously described (Villanueva et al., 2015). The K_D for each substrate was calculated with nonlinear regression in Prism software using the “One site – Specific binding” option with least squares fit.

5.2.9 Construction of cloned repressor strains

The immunity repressors from several Cluster A phages were cloned into the integrating vector pMH94 (Lee et al., 1991). The ~ 1-1.5 kb locus consisting of *rep*, its promoter, and part of the flanking upstream and downstream genes was amplified by PCR in phages L5 (coordinates 44,037-45,330 using primers oTM194, oTM195), StarStuff (coordinates 45,039-46,286 using

primers oTM196, oTM197), Et2Brutus (coordinates 44,069-45,220 using primers oTM190, oTM191), Trixie (coordinates 45,266-46,542 using primers oTM198, oTM199), Gladiator (coordinates 43,468-44,632 using primers oTM192, oTM193), and Bxb1 (coordinates 43,962-45,171 using primers oTM188, oTM189). Primers contained partial homology to pMH94 flanking the *Xba*I site. Amplicons were purified with the Nucleospin PCR Cleanup Kit, and pMH94 was linearized with *Xba*I and purified with the Nucleospin Gel Cleanup Kit. The linearized vector and amplicon were ligated using Gibson assembly (Gibson et al., 2009) and transformed into NEB5 α cells. The following plasmids were constructed: pTM75 (*L5 rep*), pTM36 (*StarStuff rep*), pTM33 (*Et2Brutus rep*), pTM38 (*Trixie rep*), pTM34 (*Gladiator rep*), and pTM32 (*Bxb1 rep*). Sequence-verified constructs were transformed into electrocompetent *M. smegmatis* mc²155. Positive transformants were selected using LB medium supplemented with kanamycin and clonally purified.

The immunity repressors from DaVinci and phiTM46 were cloned into the extrachromosomal multicopy vector pJV44 (Pham et al., 2007). The loci were amplified by PCR using primers containing *Xba*I and *Hind*III sites. For constructs containing only *rep*, analogous to the integrated repressor constructs described above, segments from DaVinci (coordinates 42,748-43,932) and phiTM46 (coordinates 42,748-43,933) were amplified using primers oTM257 and oTM265. For constructs containing the extended repressor locus (from *rep* to gene 73), segments from DaVinci (coordinates 41,377-43,932) and phiTM46 (coordinates 41,377-43,933) were amplified using primers oTM257 and oTM258. pJV44 and the amplicons were digested with *Xba*I and *Hind*III, gel-purified and cleaned up using the Nucleospin Gel Extraction Kit, ligated with T4 DNA Ligase, and transformed into NEB5 α cells. Since this cloning strategy removes the *hsp60* promoter in pJV44, a re-ligated vector backbone that lacks the *hsp60* promoter was

constructed as an empty vector control using a PCR amplicon generated from self-annealing primers oTM266 and oTM267. The following plasmids were constructed: pTM44 (empty vector), pTM48 (DaVinci *rep*), pTM53 (phiTM46 *rep*), pTM51 (DaVinci *rep*-73), pTM54 (phiTM46 *rep*-73). The repressor gene was deleted from pTM54 using the Q5 Site-Directed Mutagenesis Kit (NEB) using primers oTM290 and oTM291 according to the manufacturer's protocol to construct pTM58 (phiTM46 Δ *rep*-73). Sequence-verified plasmid constructs were transformed into electrocompetent mc²155pMH94 and mc²155pTM34 cells. Positive transformants were selected using LB medium supplemented with kanamycin and gentamicin and subsequently clonally purified. Plasmid constructs were also transformed into electrocompetent mc²155, mc²155(DaVinci), and mc²155(Gladiator), and positive transformants were selected using Middlebrook 7H10 medium supplemented with gentamicin and clonally purified.

5.2.10 Engineering L5 derivatives with C-terminally tagged repressors

Rep_{L5} was C-terminally tagged *in vivo* with either a 27 bp HA sequence (phiTM1: TACCCATACGACGTCCCAGACTACGCT) or a 24 bp FLAG sequence (phiTM6: GACTACAAGGACGACGATGACAAG)(Gordon et al., 2008) using recombineering with an L5 lysogen, similar to previous reports (Marinelli et al., 2008). The FLAG oligo (oTM51) and HA oligo (oTM52) were PCR amplified using primers oTM53 and oTM54 to create ~ 200 bp recombineering substrates that overlap the 3' end of gene 71 and that contain the tag sequence. Amplicons were purified using the GeneJet PCR Purification Kit, and the DNA was co-transformed with pJV44 into electrocompetent mc²155(L5)pJV53, as previously described (Marinelli et al., 2008). Successful pJV44 transformants were selected on Middlebrook 7H10

plates supplemented with gentamicin, and successful tag recombinants were identified by PCR. Positive recombinants were clonally purified and patched onto a lawn of mc²155, and spontaneously released phage were picked and plaque purified. From one of the HA-tagged recombinant phage picks, a spontaneous mutation was acquired (phiTM4).

5.2.11 Western blot analysis of FLAG-tagged L5 repressor

Cultures of *M. smegmatis* mc²155, or lysogens of L5 and phiTM6, were grown in Middlebrook 7H9 medium supplemented with Tween. When cultures reached an OD_{600nm} ~ 0.6-0.8, 1 ml cells were transferred to a new tube and pelleted with a brief spin in a microcentrifuge. The medium was aspirated, and the cell pellet was resuspended in Lysis Buffer (50 mM Tris pH 8.0, 300 mM NaCl, 10% glycerol) supplemented with PMSF, and 200 µl was sonicated for 10 cycles (consisting of 10 s sonication and 30 s chilling on ice). The sample was diluted with Protein Sample Buffer and run on a 15% SDS polyacrylamide gel in TGS Buffer at 100V for 2 hr. The electrophoresed samples were transferred to PVDF membrane overnight in Transfer Buffer. The membrane was blocked with TBS (20 mM Tris pH 7.4, 150 mM NaCl) + 5% milk for 2 h, incubated with 1:500 α-FLAG antibody (Cell Signaling #2368, rabbit polyclonal) in TBS + 5% BSA for 4 h, incubated with 10 ml TBS + 5% milk + 2 µl goat anti-rabbit secondary alkaline phosphatase-conjugated antibody (Life Technologies #T2191) for 1 h, and finally incubated in 16 ml dH₂O + 2 ml BCIP + 2 ml NBT (Invitrogen #002209) for 15 min. The gel was air dried and imaged. A second identical PVDF membrane was prepared, stained with Ponceau for 5 min, and imaged to assess sample loading.

5.2.12 Superinfection immunity assays

Fresh 10-fold serial dilutions of each phage lysate were generated using Phage Buffer (10 mM Tris pH 7.5, 10 mM MgSO₄, 68 mM NaCl, 1 mM CaCl₂), and 3 µl of each dilution were spotted onto a top agar layer of the indicated strain. For immunity tests involving lysogens, strains were plated in Middlebrook 7H9 top agar on Middlebrook 7H10 medium, and lysates were always spotted on an accompanying wild type mc²155 for reference. For immunity tests involving strains carrying pMH94-derived cloned repressor constructs, strains were plated in Middlebrook 7H9 + kanamycin top agar on Mycobacteria 7H11 + kanamycin medium. For immunity tests involving strains carrying pJV44-derived cloned repressor constructs, strains were plated in Middlebrook 7H9 + gentamicin top agar on Middlebrook 7H10 + gentamicin medium. For immunity tests involving strains carrying pMH94-derived constructs and pJV44-derived constructs, strains were plated in Middlebrook 7H9 + kanamycin + gentamicin top agar on Mycobacteria 7H11 + kanamycin + gentamicin medium. Lysates were always spotted on an accompanying non-lysogen or empty vector control strain (mc²155, mc²155pMH94, mc²155pTM44, or mc²155pMH94pTM44) for reference. Plates were incubated at 37°C for 3 days and photographed with ImageLab using a 1.5-2.0 s exposure. Individual assays were quantitatively scored by comparing the qualitative infection phenotypes of the phage on the strain of interest to the control strain, including efficiency of plating, turbidity, the presence of plaques, and plaque size (Table 5-2). Results were processed in R using custom scripts. More than 3,000 immunity assays were performed and manually scored, representing 1,050 unique comparisons, 239 reciprocal comparisons, and 164 lysogen-CRS paired comparisons (Supplementary Table 5-1).

Table 5-2. Infection scoring strategy.

Score	Description of phenotype
0	No spots of lysis or plaques
1	Spots of lysis at highest 1-2 titers, but no plaques
2	Challenging phage produces plaques with an efficiency of plating of less than $\sim 10^{-3}$ to 10^{-4} ; OR spots of lysis at highest 3 titers, but no plaques
3	Challenging phage produces plaques with an efficiency of plating from 10^{-1} to 10^{-3} ; OR spots of lysis at highest 4-5 titers but no plaques
4	Challenging phage produces plaques with an efficiency of plating of 1, but spots/plaques exhibit increased turbidity OR reduced size compared to infection of mc ² 155
5	Challenging phage produces plaques with an efficiency of plating of 1 and there are no phenotypic differences compared to infection of mc ² 155
6	Challenging phage produces plaques with an efficiency of plating of 1, but spots/plaques exhibit reduced turbidity OR increased size compared to infection of mc ² 155

5.2.13 Isolation of defense escape mutants

Mutant phage able to escape prophage or cloned repressor defense were isolated by picking plaques from immunity assays in which the challenging phage exhibits substantial reduction in efficiency of plating, purifying at least twice on mc²155, and confirming ability to infect the original strain. DNA was extracted from both the DEM and parent phage lysates and sequenced as previously described (Dedrick et al., 2017a). Dan Russell and Rebecca Garlena sequenced the DNA and assembled the genomes. Mutations were identified by whole genome alignment. In some cases, the parent phage genome contained one or more mutations relative to the published sequence. Only mutations that are present in the DEM compared to the parent are reported.

5.2.14 Genomic analysis of D29 and its relatives

Nucleotide sequences of D29 and its relatives were aligned using ProgressiveMauve (Darling et al., 2010), implemented through the Mauve graphical user interface (version snapshot_2015-02-25 build 0) with the following settings: default seed weight, no seed families,

LCBs determined, genomes were assumed to be collinear, with full alignment, with iterative refinement, and sum-of-pairs LCB scoring. Alignment of six Subcluster A2 genomes was performed for computing phylogenetic whole genome distance (Figure 5-2). D29 and its three nearest relatives were aligned separately for identification of single point mutations (“SNPs”) and alignment gaps that were further analyzed in Count (Csuros, 2010), Excel, and R. The phylogenetic tree was constructed on ProgressiveMauve-aligned whole genome nucleotide sequences using maximum likelihood implemented in SeaView (Gouy et al., 2010) with default settings. Trees were edited using Evolview (Zhang et al., 2012a). Alignment gaps identified by Mauve were manually examined and a sequence gap presence/absence table was constructed. Alignment gaps were mapped to branches in the D29 sub-clade tree from Figure 5-2 in Count using Dollo parsimony, which predicts whether the gap was due to an insertion or deletion event assuming a single sequence gain event with unlimited sequence deletion events.

5.2.15 Construction of strains carrying repressor-controlled P_{left} locus

To test the impact of the highly expressed P_{left} locus, plasmid constructs containing this locus under control of the thermo-inducible L5 repressor were constructed. However, creating these constructs was not straightforward and proceeded in several steps. First, the thermo-inducible repressor was cloned into the integrating plasmid pMH94, which contains two *SalI* restriction sites flanking the L5 integration cassette. The repressor locus was PCR amplified from mc²155(L5c^{ts43}), a thermo-inducible L5 lysogen (Donnelly-Wu et al., 1993), at coordinates 44,312-45,192 using Platinum Taq HiFi polymerase and 50 nt primers (oTM91 and oTM92)(Appendix B). These primers contained 25 nt at the 5' end with homology to pMH94 flanking one of the two *SalI* cut sites. The 881 nt amplicon was purified using the Nucleospin

PCR Cleanup Kit. Gibson assembly was performed according to the manufacturer's protocol to ligate together the two linearized pMH94 fragments and the repressor amplicon to construct the repressor plasmid (pTM29 and pTM31). Reactions were transformed into *E. coli*, selected with kanamycin, and verified by restriction digestion, PCR, and sequencing.

Next, the P_{left} locus was cloned into the repressor construct. The P_{left} locus was PCR amplified from the thermo-inducible L5 lysogen strain at coordinates 50,397-51,773 using Platinum Taq HiFi polymerase and 40-41 nt primers (oTM100, oTM101)(Appendix B). These primers contain 15-16 nt at the 5' end that contain *SalI* restriction sites. The 1,377 nt amplicon was purified using Nucleospin PCR Cleanup Kit and digested with *SalI*. pTM29 was digested at the remaining *SalI* restriction site and treated with CIP. The linearized pTM29 and digested amplicon were gel-purified, purified with the Nucleospin Gel Extraction Kit, ligated together, transformed into *E. coli*, and selected with kanamycin. Increased recovery efficiency of transformants occurred at lower growth temperatures of 21°C to 30°C, compared to standard 37°C. Positive transformants were verified by digestion, PCR, and sequencing, and they exhibit a mucoidy phenotype. Several constructs were sequenced, including pTM8, pTM9, pTM10, pTM11, pTM12, pTM14. No constructs contained the complete wild type P_{left} locus sequence, but instead contained several types of mutations, including single nucleotide mutations, single nucleotide deletions or large deletions that appear to have been facilitated by the frequent 13 bp stopoperator sequences present throughout this locus.

Last, the repressor-only and repressor-controlled P_{left} locus constructs were transformed into electrocompetent *M. smegmatis* mc²155 cells and selected with kanamycin. As with the *E. coli* strains, increased recovery efficiency of transformants occurred at lowered growth

temperatures of 21°C to 30°C. Positively transformed colonies were noticeably smaller than wild type colonies. Transformants were purified and verified by PCR.

5.2.16 P_{left} toxicity test

To test the impact of the highly expressed locus downstream of P_{left} , temperature sensitive growth assays were performed. Liquid cultures of *M. smegmatis* strains were grown from single colonies in 3 ml Middlebrook 7H9 + 4 µg/ml kanamycin + 0.05% Tween 80 at 30°C, shaking at 250 rpm. Cultures were grown for 5 days and diluted to $\text{OD}_{600\text{nm}} = 0.5$. Ten-fold serial dilutions were made for each culture, and 3 µl of the 10^0 to 10^{-6} dilutions were spotted onto two sets of Middlebrook 7H10 + kanamycin plates. One set was incubated at 30°C or 44°C. After four days of growth, images of plates were taken using ImageLab.

5.2.17 Data analysis

Infection data were analyzed and visualized in RStudio (<https://www.rstudio.com>) using custom scripts with the *psych*, *reshape2*, and *stringdist* packages. The scripts are available on Github. More than 65% of unique comparisons were measured with two or more replicate assays, and replicate infection scores were averaged. Although the infection score can vary between replicates, more than 80% of comparisons with two or more replicates exhibit a range of infection scores smaller than 2. All R^2 correlations between genetic elements and immunity phenotypes were determined with linear regression only using intra-L5 clade comparisons (unless otherwise indicated) and using the *lm* function.

5.3 RESULTS

5.3.1 Characterization of the Cluster A immunity system

All Cluster A phages exhibit similar genomic architectures: the left arm contains structural and assembly genes, the right arm contains genes associated with lytic growth such as DNA replication, and the genome center contains prophage inheritance genes such as integration or partitioning systems (Figures 1-2, 4-1A)(Pope et al., 2011b). The immunity repressor can be readily identified at syntenic positions, and the right genome terminus harbors the early lytic promoter, P_{left} . Despite the conserved synteny, these phages are genetically diverse and have been further subdivided into 19 subclusters (Figure 5-1A). Phages from distinct subclusters, such as Bxb1 (Subcluster A1), L5 (Subcluster A2), and Peaches (Subcluster A4) exhibit highly divergent repressors and stopoperator motifs and are heteroimmune (Pope et al., 2011b). However, the genetic diversity within and between subclusters is not homogenous, and there is a clade of over 100 phages representing ten subclusters that are more closely related to L5 than to Bxb1 or Peaches, and they exhibit a spectrum of genetic diversity based on their gene content and nucleotide sequence (Figure 5-1A, B).

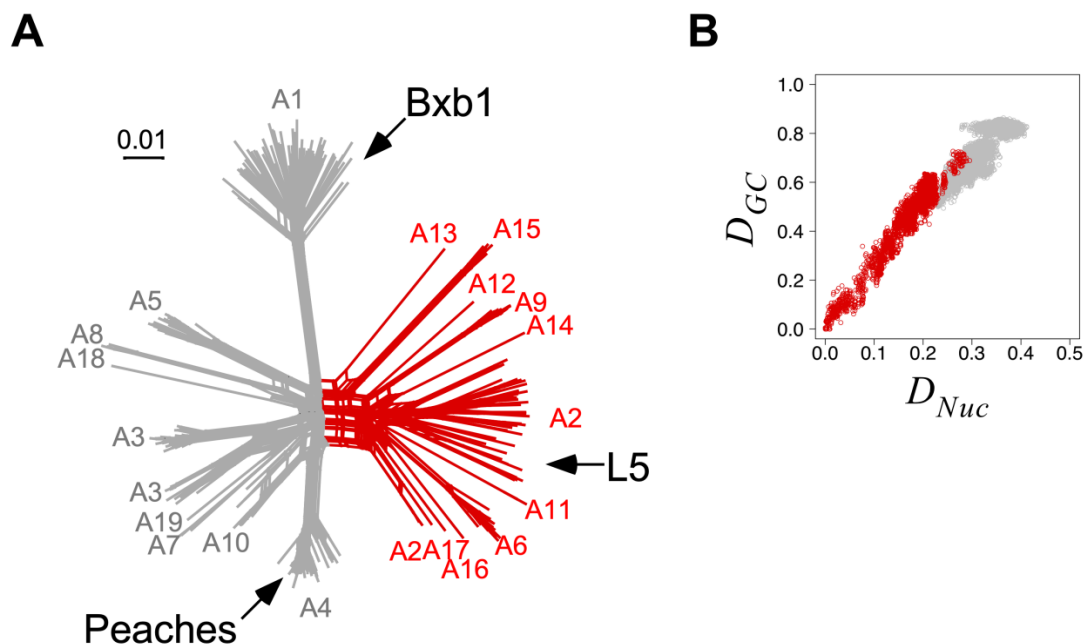


Figure 5-1. Immunity system of phages in the L5 clade exhibits a genetic spectrum.

(A) Phylogenetic network of 311 Cluster A phages based on gene content using Splitstree. Groups of taxa are labeled with their subcluster designation, several phages are labeled for reference and a clade of phages representing ten subclusters that are more closely related to L5 than others are highlighted in red. (B) Scatter plot comparing whole genome nucleotide (D_{Nuc}) and gene content (D_{GC}) distances involving one Cluster A *Mycobacterium* phage within the L5 clade with another Cluster A *Mycobacterium* phage within (red) or without (grey) the L5 clade. Number of *Mycobacterium* phages in L5 clade = 87.

This diversity can be illustrated with phages L5 and D29. Historically, L5 and D29 were two of the first *Mycobacterium* phages to be sequenced and characterized (Ford et al., 1998; Hatfull and Sarkis, 1993). They are more closely related to each other than to Bxb1 and are grouped in Subcluster A2 (Figure 5-2A). D29 is able to superinfect a Bxb1 lysogen but not an L5 lysogen (Donnelly-Wu et al., 1993; Mediavilla et al., 2000). D29 nevertheless exhibits substantial sequence divergence from L5 (they exhibit 84% sequence identity across 75% of their genomes), and it is a lytic mutant, in which a portion of the right end of the genome downstream of P_{left} , including the 5' end of *rep*, has been deleted (Ford et al., 1998). As a result of this mutation, D29 lysogens are not able to be generated, preventing further investigation into the D29 immunity system. However, there are now several close temperate relatives to both D29 and L5 that have been isolated (Figure 5-2A, B). The closest relative to L5 is now Serenity, exhibiting 91% sequence identity across 87% of their genomes. StarStuff, Pomar16, and Kerberos each exhibit 98% sequence identity with D29, and they form a distinct phylogenetic sub-clade from L5 and Serenity (Figure 5-2B). Unlike D29 though, none of these phages exhibit the same deletion in the right arm of their genomes, and they encode a complete immunity repressor. A StarStuff lysogen can be generated, and it is homoimmune to L5 and heteroimmune to Bxb1 (Figure 5-2C). However, a StarStuff lysogen exhibits stronger immunity to D29 than an L5 lysogen does. Although D29 does not form plaques on an L5 lysogen, it does form clearings at the highest titers. The genetic differences between D29 and L5 may be sufficient enough to enable D29 to overcome L5 immunity at high multiplicities of infection, similar to previous reports (Donnelly-Wu et al., 1993). Therefore, the 100 phages within the L5 clade may harbor mesotypic immunity systems that exhibit an entire spectrum of diversity. To investigate this, I characterized and compared several genetic elements of their immunity systems.

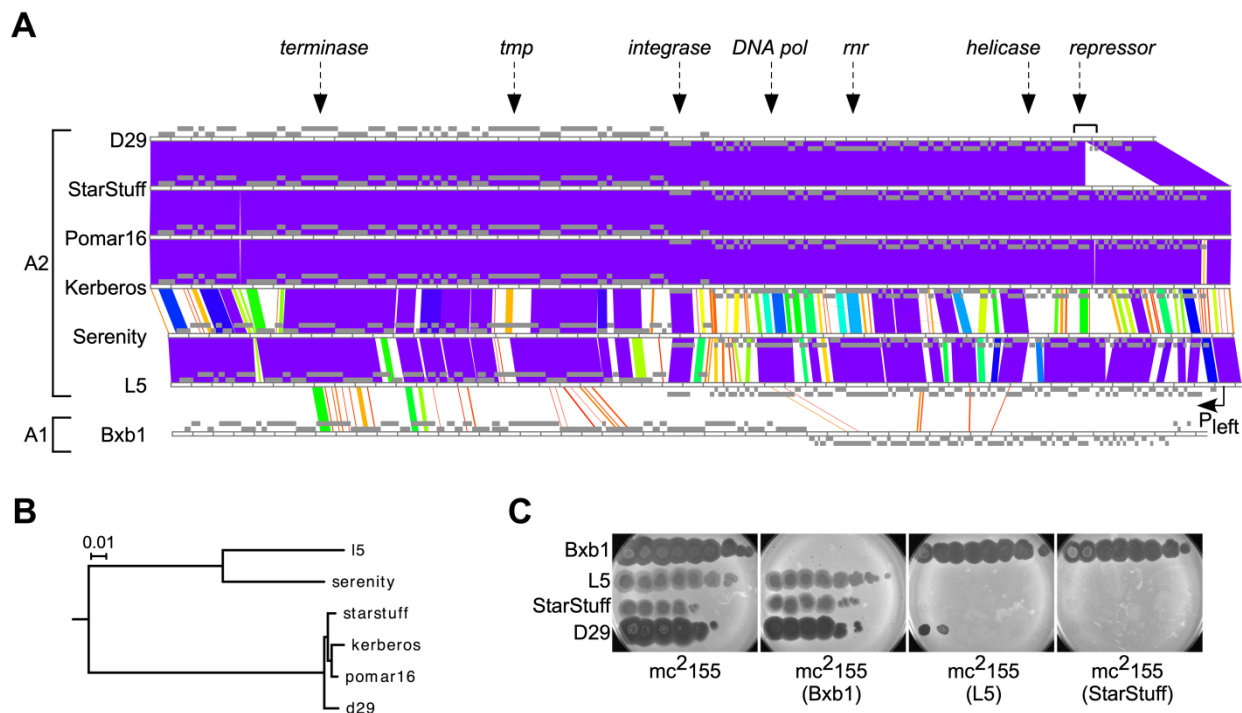


Figure 5-2. Genomic relationship of D29 to other Cluster A relatives.

(A) Pairwise BLAST-based whole genome alignments of D29 and several Cluster A *Mycobacterium* phages with subclusters indicated. Grey boxes above and below the ruler in each map represent genes transcribed on the top and bottom strands, respectively. The color spectrum between genomes represents pairwise nucleotide similarity, from high similarity (violet) to no detectable similarity (white). Maps were generated using Phamerator. The positions of several genes highly conserved in Cluster A phages are indicated by reference (*tmp* = tape measure protein; *pol* = DNA polymerase; *rnr* = ribonucleotide reductase). **(B)** Phylogenetic tree constructed from whole genome alignment of phages in panel A using Mauve. **(C)** Superinfection immunity assays using 10-fold serial dilutions of phages Bxb1, L5, StarStuff, and D29 against mc²155 and lysogens (Bxb1, L5, and StarStuff). Figure adapted from (Dedrick et al., 2017b).

Immunity repressors in the L5 clade are similar in size and exhibit a genetic spectrum that correlates with whole genome gene content distances (Figure 5-3A, B). As seen with L5, Bxb1, and Peaches, a set of stoperator sites can be identified in each genome (Pope et al., 2011b) (see Materials and Methods). Similar amounts of stoperators are present in each genome and they are predominantly oriented in the direction of transcription (Figure 5-4A, B)(Brown et al., 1997; Jain and Hatfull, 2000). Sequence motifs representing each genome's cognate stoperators are similar, but not identical, to each other, and they exhibit a genetic spectrum that also correlates with whole genome gene content distances and repressor genetic distances (Figure 5-4C, D).

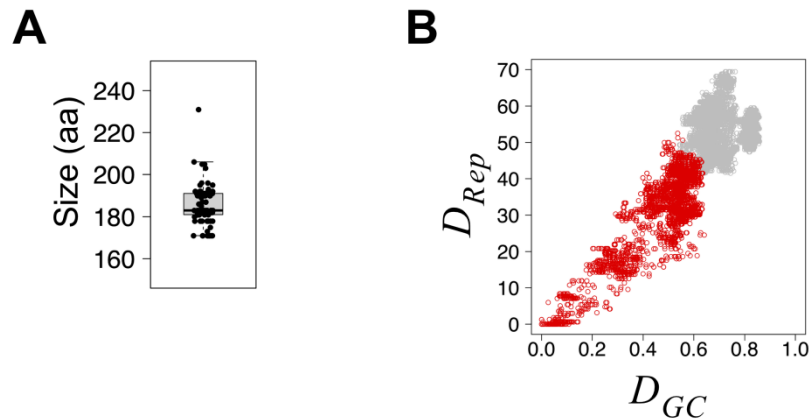


Figure 5-3. Characterization of Cluster A immunity repressors.

(A) Box plot of immunity repressor amino acid sizes from 82 L5 clade phages. **(B)** Scatter plot comparing pairwise whole genome gene content (D_{GC}) and Rep (D_{Rep}) genetic distances between Cluster A phages as in Figure 5-1B.

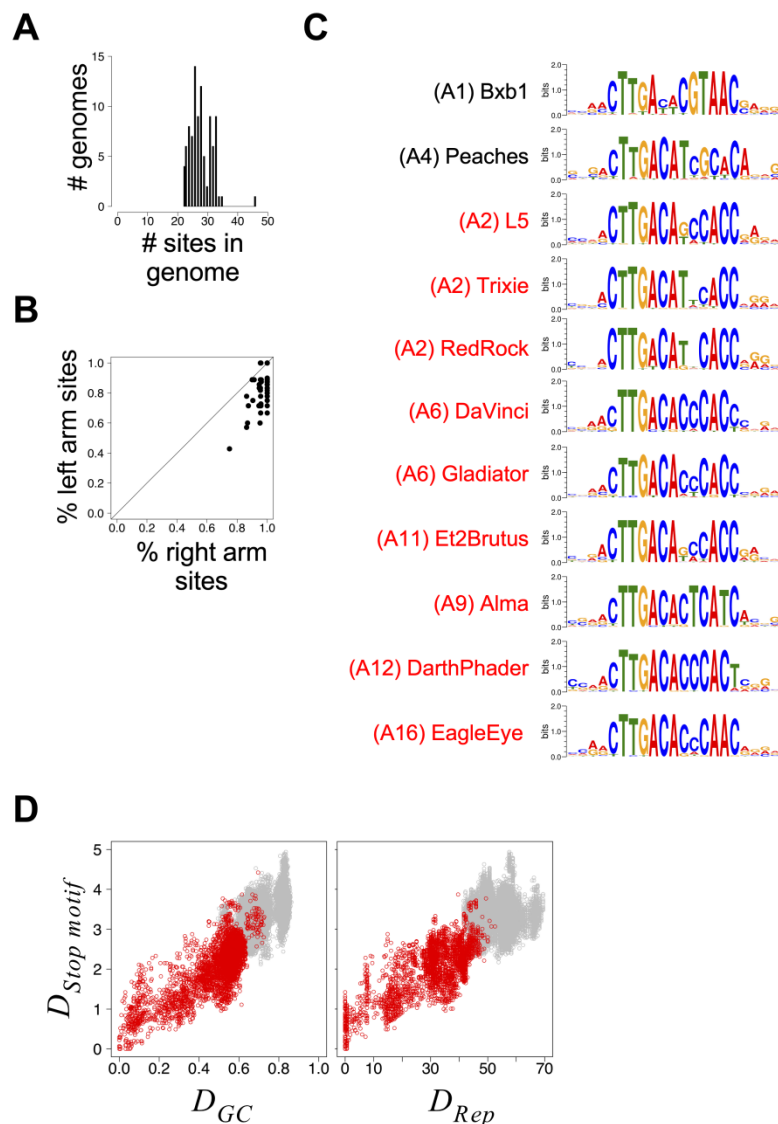


Figure 5-4. Characterization of Cluster A stoperators.

(A) Histogram reflecting the number of L5 clade phage genomes that contain the indicated amount of stoperator sites. (B) Scatter plot comparing the percentage of stoperator sites positioned in the right or left arm of the genome that are oriented in the direction of transcription. The $y = x$ line is plotted for reference. (C) Alignment of sequence motif logos representing predicted stoperator sites for several phages within the L5 clade, with their subcluster indicated. (D) Scatter plot comparing whole genome gene content (D_{GC}) or Rep (D_{Rep}) genetic distances with stoperator motif distances ($D_{Stop\ motif}$) between phages as in Figure 5-1B.

Since L5 clade phages are genetically diverse, gene regulation during lysogeny may or may not be similar, so I compared gene expression profiles to stoperator sites in several phages.

In addition to the genome-wide expression profiles described for extrachromosomal L5 clade phages in Chapter 4 (representing Subclusters A2, A9, and A16), I performed strand-specific RNAseq on Et2Brutus (Subcluster A11), Gladiator (Subcluster A6), and Trixie (Subcluster A2) lysogens. Also, Rebekah Dedrick and I performed strand-specific RNAseq on L5, StarStuff, and D29 (Subcluster A2) at several time points after infection, ranging from 15 min to 2.5 h, as well as on L5 and StarStuff lysogens. Dan Russell and Rebecca Garlena sequenced the RNA.

Similar to gene expression patterns in the prophages RedRock, EagleEye, Alma, and Pioneer described in Chapter 4, very few genes are expressed in Et2Brutus, Gladiator, and Trixie prophages during lysogeny (Figure 5-5A). Similar results are observed for L5 and StarStuff (data not shown). The strongest gene expressed is *rep*, and weak expression is observed at P_{left} and the prophage inheritance loci. Over 20% of stoperators are located within 1.5 kb of the right genome termini in proximity to P_{left} . Expression at the *rep* locus is very similar between prophages, initiating in the upstream intergenic region where promoters in L5 have been mapped (Nesbit et al., 1995) and substantially decreasing within ~ 1 kb downstream near the end of the *cas4-family* gene (Figure 5-5B). Many phages in this clade contain stoperators within this promoter as well as near the 3' end of the *cas4-family* gene suggesting conserved regulatory strategies are utilized at this locus. At P_{left} , the weak expression observed during lysogeny or the strong expression observed during lytic growth consistently begins within a cluster of stoperators, one of which functions as the empirically-determined operator in L5 (Figure 5-5C)(Brown et al., 1997; Nesbit et al., 1995). Expression extends ~ 100-150 bp downstream across a locus devoid of stoperators, terminating prior to the first coding sequence. Most phages in this clade have similarly positioned stoperators at this locus, suggesting conserved regulatory strategies are utilized at P_{left} as well.

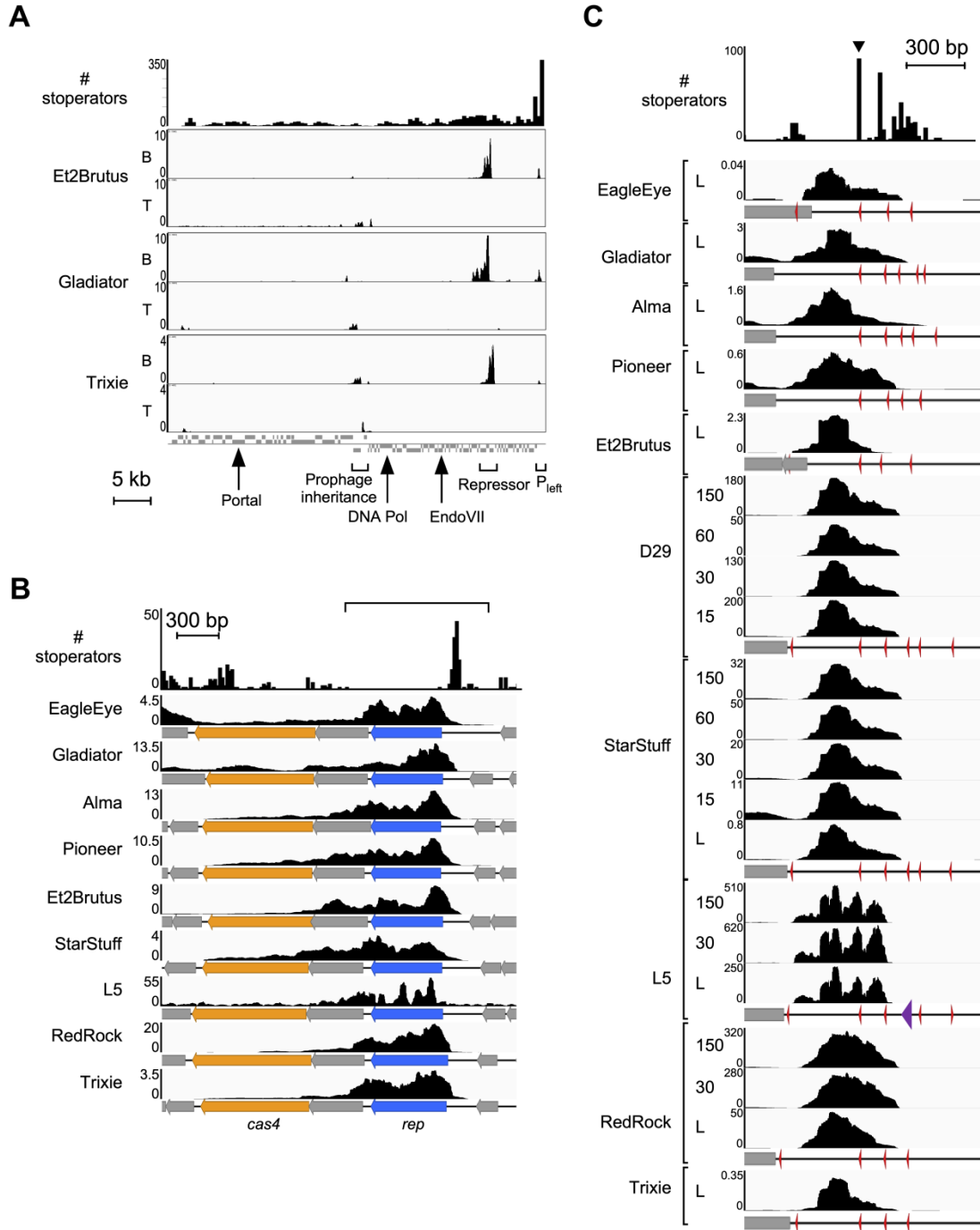


Figure 5-5. Expression patterns of Cluster A phages during lysogeny and infection.

(A) Strand-specific RNAseq expression profiles for top (T) and bottom (B) strands of Et2Brutus (Subcluster A11), Gladiator (Subcluster A6), and Trixie (Subcluster A2) prophages during lysogeny. Specific loci of interest are highlighted. Pol: Polymerase; EndoVII: Endonuclease VII. Prophage inheritance: integrase or *parABS* locus. Histogram above expression profiles reflects the genomic distribution of stoperators identified in all L5 clade phages relative to the right end of the genome. (B) Enlarged view of bottom strand expression profiles across the

repressor locus for phages from panel A as well as for several previously reported Cluster A phages, with *rep* (blue) and the *cas4*-like (orange) genes indicated. Histogram above the expression profiles reflects the distribution of stoperator sites in all L5 clade phages relative to the 3' end of the repressor gene. The region cloned from several phages to test for repressor-mediated immunity is indicated above with brackets. (C) Enlarged view of bottom strand expression profiles at the P_{left} locus for phages from panel A as well as for several Cluster A phages during lysogeny (L) and lytic growth (15, 30, 60, or 150 min post infection). Grey boxes: genes; red arrowheads: predicted stoperators; purple arrowhead: empirically identified L5 operator. Histogram above the expression profiles reflects the distribution of stoperator sites in all L5 clade phages relative to the stoperator used (black arrowhead) for manual alignment of genomes below.

Although it is not known whether Cluster A phages utilize all stoperators in the genome, their sequence diversity in L5 and Bxb1 correlates with empirically-determined *in vitro* binding specificity for L5 and Bxb1 repressors (Brown et al., 1997; Jain and Hatfull, 2000). To examine whether predicted stoperators in other Cluster A phage genomes also reflect general binding specificity, I tested binding specificity of the Trixie (Subcluster A2) Rep for bioinformatically identified cognate and non-cognate stoperators. I cloned, overexpressed, and purified Rep_{Trixie} (Figure 5-6A). I tested the repressor's affinity for syntenically positioned stoperators from phages of several subclusters (Figure 5-6B). Rep_{Trixie} exhibits strongest affinity for Trixie and RedRock (Subcluster A2) stoperators, slightly reduced affinity for a Gladiator (Subcluster A6) stoperator, and very low affinity for Alma (Subcluster A9), Rockstar (Subcluster A3), and Peaches (Subcluster A4) stoperators and a substrate with no predicted stoperator (Figure 5-6C, D, E). Furthermore, binding affinity is progressively diminished when the substrate's sequence is incrementally changed from a Trixie to a Peaches stoperator (Figure 5-6E, F). Thus, among phages in the L5 clade, bioinformatically identified stoperators generally reflect the endogenous repressor's binding affinity. The diversity in these phages may reflect an evolving immunity system that retains similar genomic architecture but is comprised of a genetic spectrum of regulatory elements.

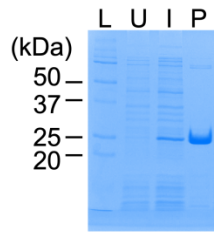
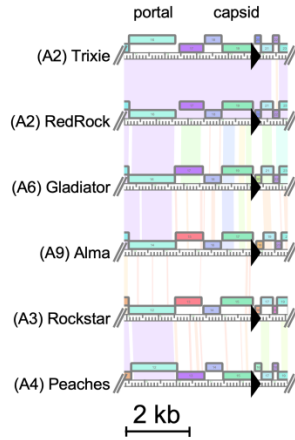
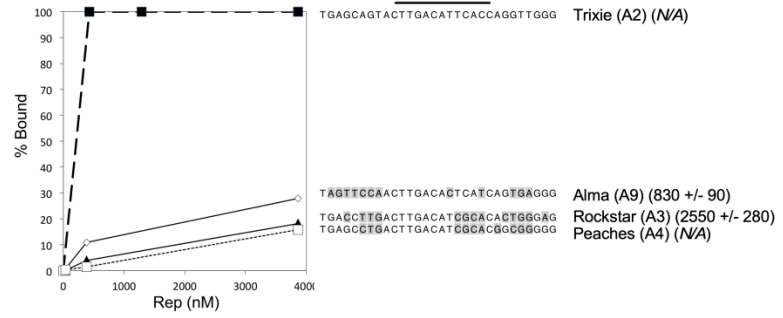
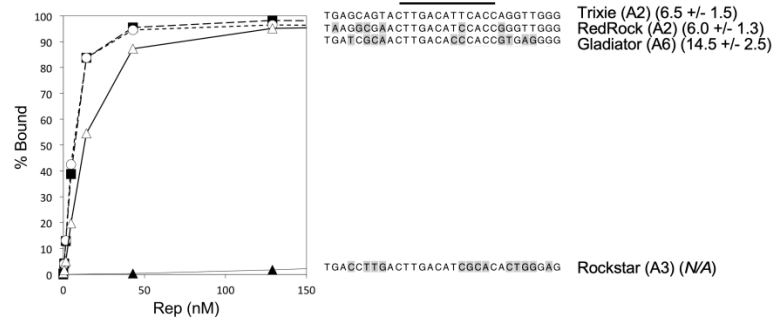
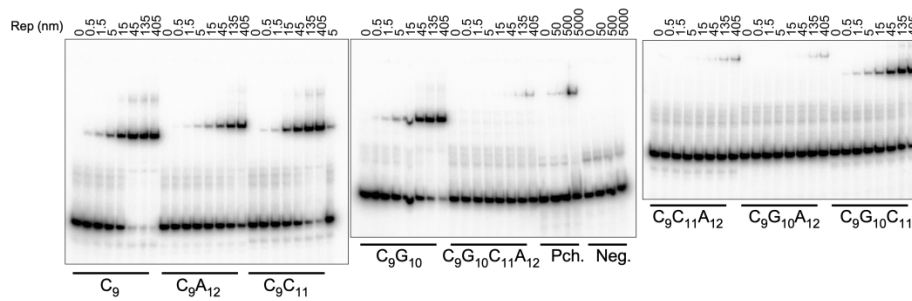
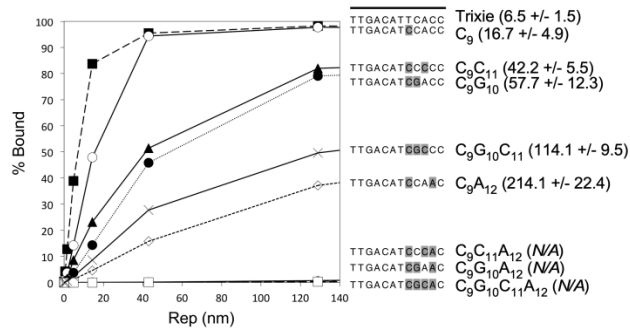
A**B****C****D****E****F**

Figure 5-6. Characterization of Trixie repressor *in vitro* binding affinity.

(A) Coomassie stained polyacrylamide gel reflecting purification of Rep_{Trixie}, which was cloned in the pET-21a vector, tagged with His, overexpressed by IPTG induction, and purified using a Nickel affinity matrix. L: ladder with specific kDa bands labeled; U: uninduced whole cell lysate; I: induced whole cell lysate; P: purified protein eluted from Nickel affinity matrix. (B) Enlarged view of the *capsid* locus from a whole genome alignment of several Cluster A phages using Phamerator. Each genome contains a predicted stoperator (black arrowhead) at the syntenic position immediately downstream of the *capsid* gene. (C-D) Quantification of electrophoretic mobility shift assays (EMSAs, not shown) in which the binding affinity of purified Rep_{Trixie} for different stoperators at the *capsid* locus was tested. Varying amounts of Rep were incubated with radiolabeled 30 bp dsDNA substrate from the syntenic *capsid* locus (right) that contained the 13 bp predicted stoperator site (indicated by black line), and the % of DNA bound by Rep was measured (left). The K_D and standard error indicated at far right. (E) The binding affinity of Rep_{Trixie} was tested for a series of 30 bp substrate containing progressive nucleotide changes between the homologous 13 bp Trixie and Peaches stoperator sites. (F) EMSA results in panel E were quantified as in panels C-D.

5.3.2 L5 clade phages exhibit mesoimmunity

To determine how an evolving immunity system impacts superinfection immunity, I selected 19 phages from 7 subclusters across the L5 clade representing varying degrees of genetic diversity based on their gene content, immunity system regulatory elements, and prophage inheritance strategies (Table 5-1). Lysogens were generated with each phage, as well as with Dreamboat (Subcluster A1) as a heterotypic Cluster A control. Superinfection immunity assays were performed against these lysogens using a variety of phages, including the parent temperate phages from which the lysogens were generated, several lytic mutants isolated from the environment, and several heterotypic Cluster A phages including Peaches (Subcluster A4), Bxb1 (Subcluster A1), and Petruchio (Subcluster A1).

In many examples of immunity, phages tend to exhibit a symmetric binary phenotype. Superinfection of a lysogen harboring a heteroimmune prophage occurs with the same efficiency of plating as infection of a non-lysogen, and superinfection of a lysogen harboring a

homoimmune prophage fails to produce plaques. The reciprocal infection test produces the same, symmetric results. However, superinfection phenotypes between L5 clade phages extend beyond these delineations (Figure 5-7). For example, although there is no phenotypic distinction between L5 infection of an EagleEye lysogen and mc²155, or between L5 infection of a StarStuff lysogen and an L5 lysogen, L5 infections of DarthPhader and ArcherNM lysogens produce intermediate phenotypes that reflect neither complete infection nor complete immunity. In contrast, Peaches superinfection phenotypes on all lysogens are nearly indistinguishable from infection of mc²155 (Figure 5-7). The intermediate phenotypes are diverse and are not simply reductions in efficiency of plating; there are changes in the size and turbidity of plaques and spots, and sometimes spot dilutions appear to fade away with no discernible individual plaques (Figures 5-7, 5-8). Some infections appear to be enhanced on lysogens compared to mc²155, in which plaques or spots are enlarged or less turbid, such as Gladiator superinfecting an Alma lysogen or DarthPhader superinfecting an ArcherNM lysogen (Figure 5-8A, D). Additionally, several reciprocal infection tests do not result in symmetric phenotypes. An L5 lysogen is sensitive to Trixie superinfection, but a Trixie lysogen completely defends against L5 superinfection (Figure 5-9A). Similarly, an EagleEye lysogen is sensitive to DaVinci superinfection, but a DaVinci lysogen completely defends against EagleEye superinfection (Figure 5-9B).

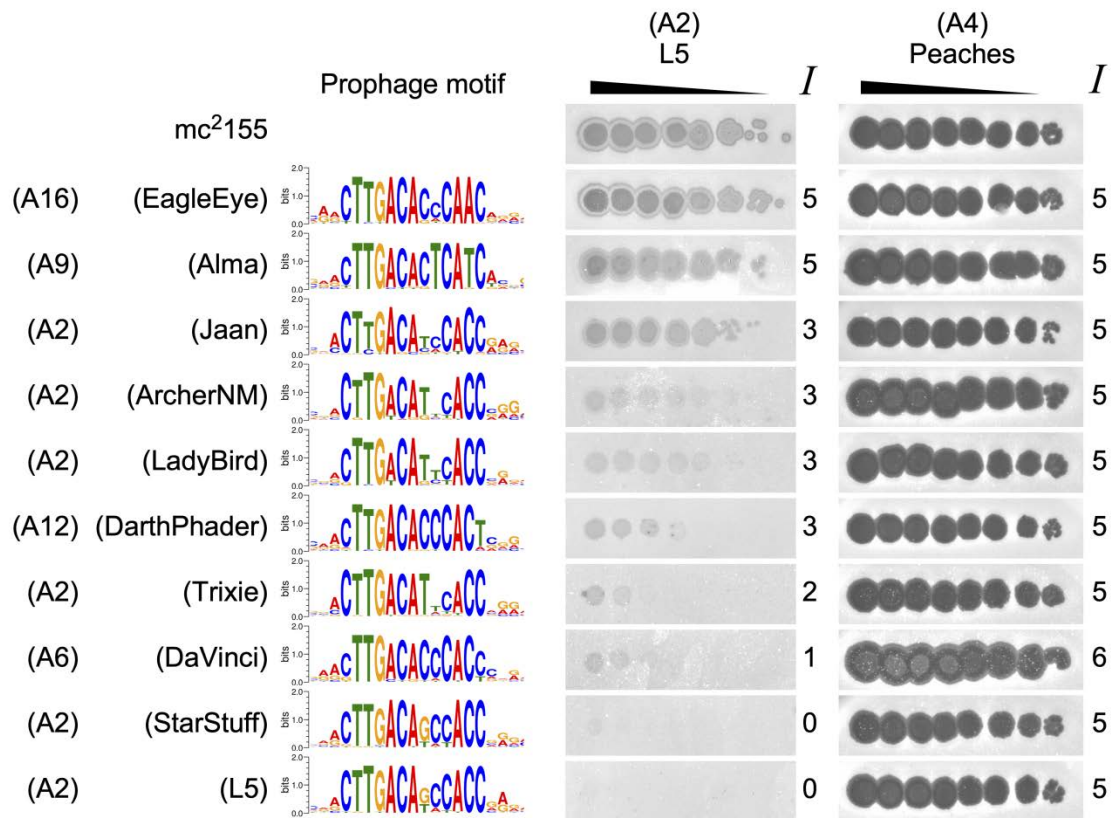


Figure 5-7. L5 exhibits a spectrum of infection phenotypes.

Representative immunity assays exhibiting infection phenotypes of L5 against *M. smegmatis* mc²155 and lysogens harboring prophages from the L5 clade. Peaches (Subcluster A4) is used as a heterotypic control. The subclusters and stoperator motifs for each prophage are indicated. Black triangles indicate 10-fold serial dilutions of phage lysate. Infection phenotypes (*I*) on lysogens are scored relative to the infection phenotype on mc²155.

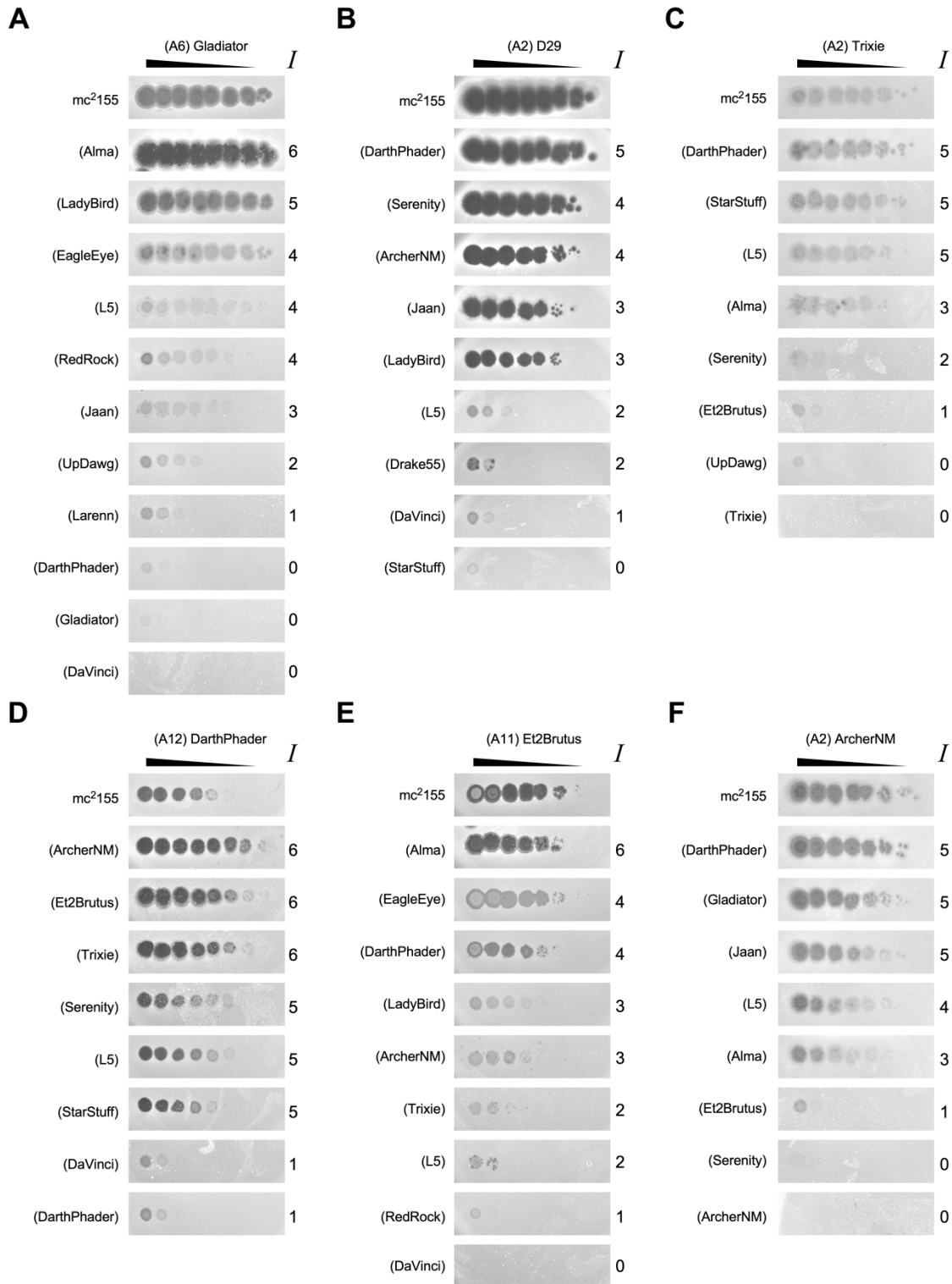


Figure 5-8. Multiple L5 clade phages exhibit a spectrum of infection phenotypes.

Representative immunity assays for several L5 clade phages in as in Figure 5-7, demonstrating incomplete infection phenotypes and their infection scores.

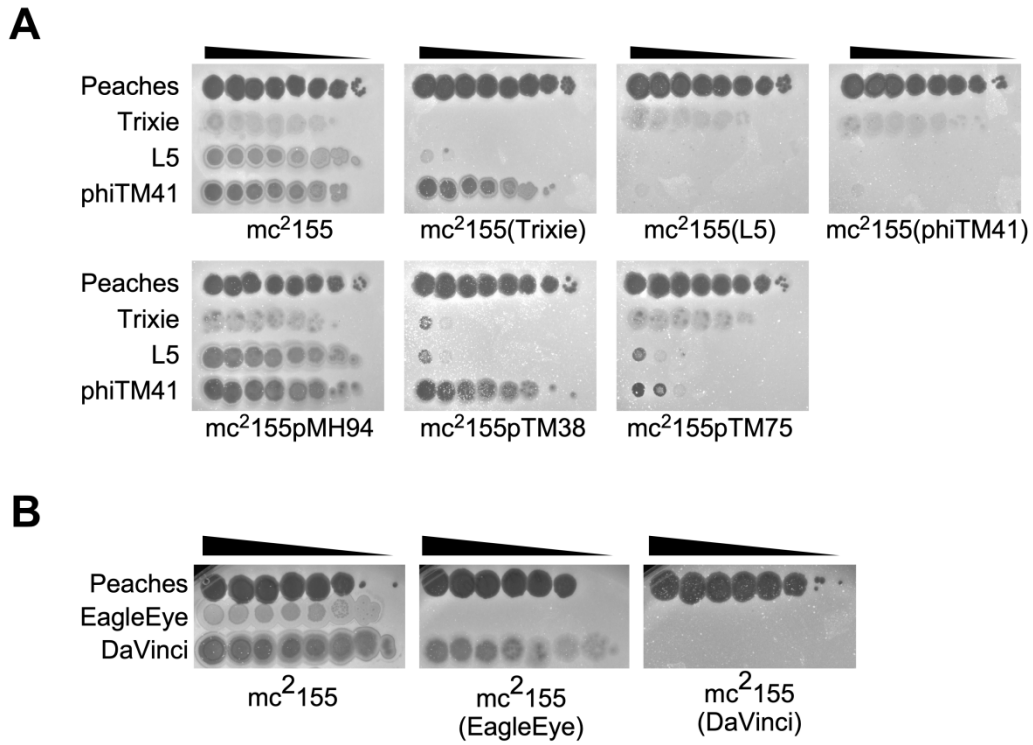


Figure 5-9. L5 clade phages exhibit asymmetric immunity.

Representative immunity assays comparing infection phenotypes of **(A)** Trixie, L5, and phiTM41 against mc²155, lysogens (Trixie, L5, and phiTM41), and CRSs (pMH94: empty vector; pTM38: Trixie; pTM75: L5), or **(B)** EagleEye and DaVinci against mc²155 and lysogens (EagleEye, DaVinci). Peaches serves as a heterotypic control.

The biological relevance of the diverse infection phenotypes is not obvious. Therefore, I scored each phage's superinfection phenotype on each lysogen relative to its infection phenotype on mc²155 and compared phenotypes across the L5 clade of phages (Figures 5-7, 5-8). Infection scores (*I*) ranged from 0 (no observable infection phenotype) to 6 (efficiency of plating of 1 and enlarged plaques, relative to infection on mc²155)(Table 5-2). As expected, no plaques are observed (*I* < 2) when L5 clade lysogens are challenged by their parent phage. Consistent with previous reports, in the majority of immunity assays involving heterotypic Cluster A phages (Peaches, Bxb1, Dreamboat, and Petruchio), superinfections produced phenotypes with no reduction in efficiency of plating, reduction of plaque size, or increase in turbidity relative to mc²155 (*I* > 4)(Figure 5-10)(Pope et al., 2011b). However, in only 55% of assays involving two different phages in the L5 clade, superinfections produced either no phenotype (*I* = 0) or a phenotype identical to infection of mc²155 (*I* = 5) (Figure 5-11A). The intermediate phenotypes (*I* = 1-4) or enhanced infection phenotype (*I* = 6) are observed with a substantial number of superinfecting phages and defending prophages, indicating the phenotypes are not caused by any particular phage or prophage (Figure 5-11A). Fewer than 50% of all inter-subcluster assays resulted in superinfection phenotypes identical to infection on mc²155 (*I* = 5), and fewer than 50% of all intra-subcluster assays resulted in no infection phenotype (*I* = 0), indicating the superinfection phenotypes do not simply correlate with subcluster designations (Figure 5-11B). Additionally, many reciprocal assays exhibit asymmetric superinfection phenotypes, such as those involving phage Jaan (Figures 5-10, 5-11C). Overall, although homoimmunity and heteroimmunity are both present within the L5 clade, delineation of distinct immunity groups is obfuscated by mesoimmunity phenotypes.

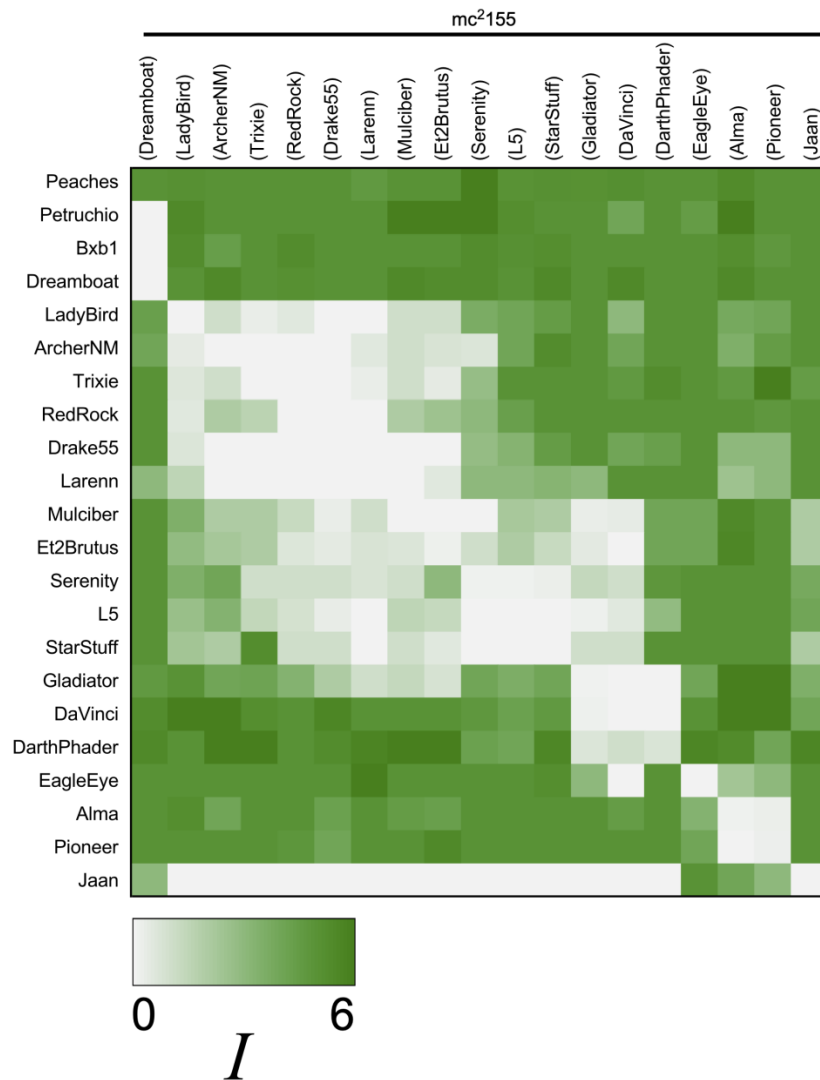


Figure 5-10. Mesoimmunity phenotypes among L5 clade phages.

Heatmap matrix of averaged infection scores of superinfecting L5 clade phages (rows) against defending L5 clade prophages (columns), where green indicates stronger infection ($I = 6$) and white indicates stronger defense ($I = 0$). Peaches, Dreamboat, Petruchio, and Bxb1 are used as heterotypic controls.

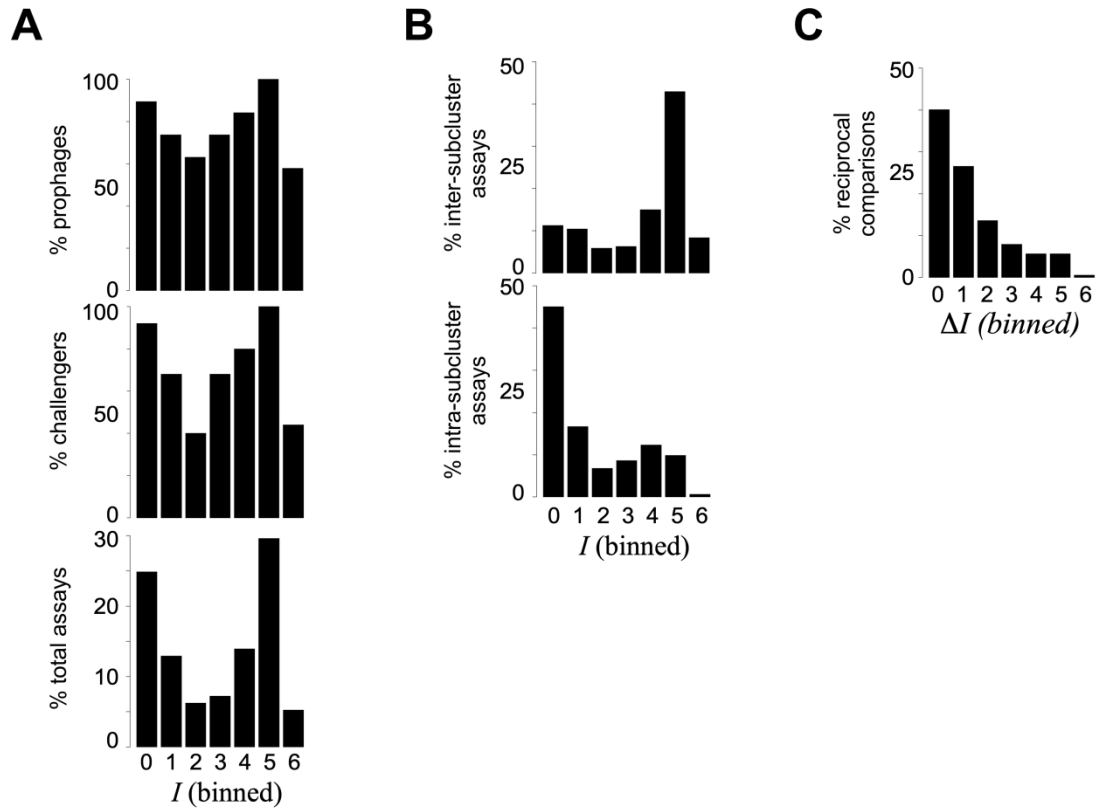


Figure 5-11. Quantification of asymmetric and incomplete immunity phenotypes.

(A) Histograms summarizing immunity assays involving L5 clade phages. Replicate infection scores for unique defending-challenging phage assays were averaged and binned, and histograms reflect the percentage of (top) defending prophages, (middle) challenging phages, and (bottom) all unique assays associated with each score. **(B)** Histograms as in panel A for (top) assays only involving phages in different subclusters, and (bottom) assays only involving phages in the same subcluster. **(C)** Histogram summarizing percentage of binned infection score difference (ΔI) between reciprocal defending-challenging phage infection tests.

5.3.3 Mesoimmunity phenotypes correlate with immunity system evolution

Although several genetic factors can impact superinfection, the mesoimmunity phenotypes are likely caused by changes in the immunity system. Many prophages express genes other than the immunity repressor to exclude phages, and many phages utilize strategies to overcome these defenses (Bondy-Denomy et al., 2016; Dedrick et al., 2017a). However, Cluster A prophages do not express many genes during lysogeny (Figures 4-5, 5-5A). Additionally, since the immunity phenotypes correlate with changes in whole genome gene content and nucleotide sequence, they are not likely caused by the gain or loss of individual genetic loci that work to defend against phages during lysogeny or overcome prophage defenses during superinfection (Figure 5-12A). Incompatibilities between integrases or partitioning systems may impact superinfection (Chapter 4), but mesoimmunity is observed among phage pairs regardless of prophage inheritance genes (Figure 5-12B, C). Instead, immunity phenotypes correlate with changes in repressors or stoperator motifs, and more so than with other highly conserved genes (Figure 5-13A, B). Additionally, pairwise correlations between phage defense or superinfection profiles decrease as stoperator motif distances increase (Figure 5-13C).

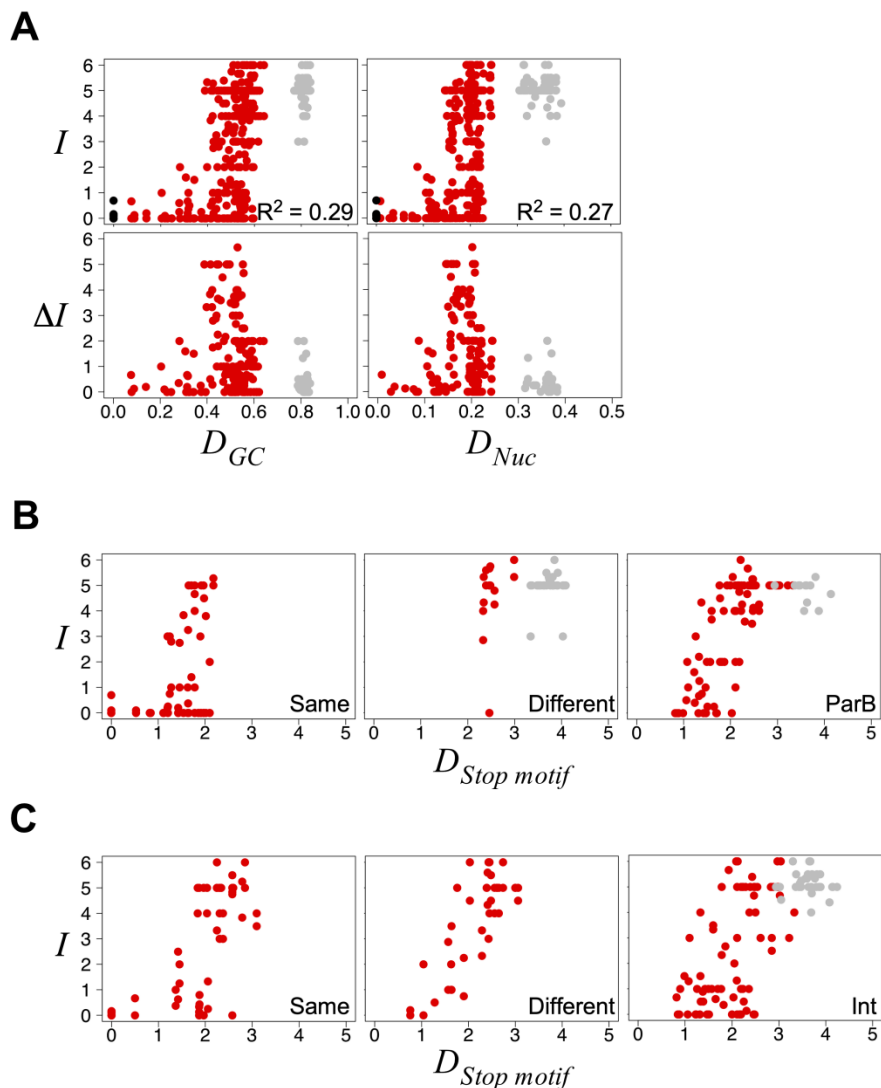


Figure 5-12. Mesoimmunity phenotypes correlate with whole genome evolution.

(A) Scatter plots involving an L5 clade phage with itself (black), another L5 clade phage (red), or a non-L5 clade phage (grey) comparing (top, $n = 423$) infection scores (I) or (bottom, $n = 185$) reciprocal infection score differences (ΔI) to whole genome gene content (D_{GC}) or nucleotide (D_{Nuc}) distances. (B) Scatter plots comparing the average infection score with the stopoperator motif distance ($D_{Stop\ motif}$) between an integrated defending prophage and (Same, $n = 65$) a challenging phage that contains an integrase gene in the same pham as the defending phage's integrase, (Different, $n = 52$) a challenging phage that contains an integrase gene in a different pham than the defending phage's integrase, or (ParB, $n = 96$) an extrachromosomal challenging phage. (C) Scatter plots comparing the infection score with the stopoperator motif distance ($D_{Stop\ motif}$) between an extrachromosomal defending prophage and (Same, $n = 58$) a challenging phage that contains a ParB gene in the same pham as the defending phage's ParB, (Different, $n = 42$) a challenging phage that contains a ParB gene in a different pham than the defending phage's ParB, or (Int, $n = 110$) an integrating challenging phage.

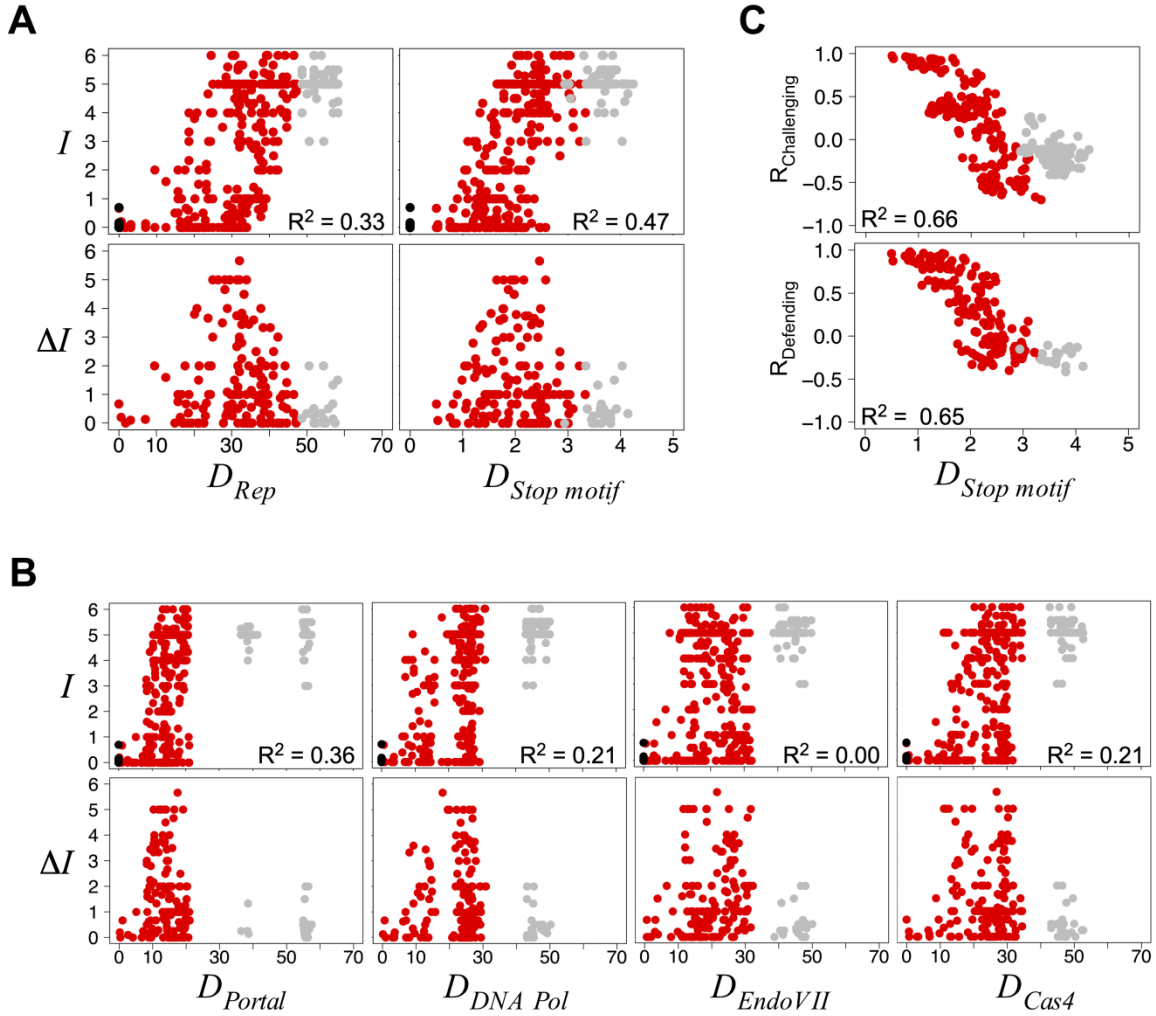


Figure 5-13. Mesoi mmunity phenotypes correlate with immunity system diversity.

(A) Scatter plots for all immunity assays involving an L5 clade phage with either itself (black), another L5 clade phage (red), or a non-L5 clade phage (grey) comparing Rep genetic distance (D_{Rep}) or the stoperator motif distance ($D_{Stop\ motif}$) with the infection phenotype (I) of individual immunity assays (top, $n = 423$) or the change in infection phenotype (ΔI) between reciprocal immunity assays (bottom, $n = 185$). The R^2 value from a linear regression of all data involving two L5 clade phages is indicated. (B) Scatter plots comparing (top, $n = 423$) the averaged infection score (I) or (bottom, $n = 185$) the infection score difference (ΔI) between reciprocal assays to the genetic distance of the highly conserved Portal, DNA Polymerase, EndoVII, and Cas4 genes. R^2 value from a linear regression is indicated. Color schemes as in Figure 5-12. (C) Scatter plots for pairs of phages comparing the stoperator motif distance ($D_{Stop\ motif}$) with (top, $n = 225$) the correlation coefficient of the two phages' superinfection profiles against the panel of lysogens ($R_{Challenging}$), or (bottom, $n = 171$) the superinfection profiles of phages against the two phages' lysogens ($R_{Defending}$).

5.3.4 Mesoimmunity is repressor-mediated

Rep_{L5} is necessary and sufficient to confer immunity against L5 superinfection (Donnelly-Wu et al., 1993). Therefore, I tested whether mesoimmunity phenotypes are observed using strains that harbor cloned immunity repressors. I cloned the ~ 1-1.5 kb homologous repressor locus of six phages (from multiple subclusters) into integrating vectors, as previously described for L5 (Donnelly-Wu et al., 1993), to create a series of cloned repressor strains (CRSs)(see Materials and Methods). Every cloned immunity repressor, such as Gladiator, confers strong homotypic defense (Figure 5-14A). CRS immunity is not always complete though, and increased infection phenotypes are sometimes observed, as previously observed for L5 (Figure 5-14A)(Donnelly-Wu et al., 1993). Additionally, a Trixie lysogen and CRS exhibit immunity profiles consistent with the presence of empirically determined Trixie binding sites in each genome (Figure 5-14B). Next, I tested superinfection of diverse phages against each CRS. The strengths of infection on CRSs and lysogens are highly correlated (Figure 5-14C). Asymmetric phenotypes are also observed with CRSs: a Trixie CRS (mc²155pTM38) defends against L5 superinfection while an L5 CRS (mc²155pTM75) does not defend against Trixie superinfection, consistent with the lysogen immunity patterns (Figure 5-9A).

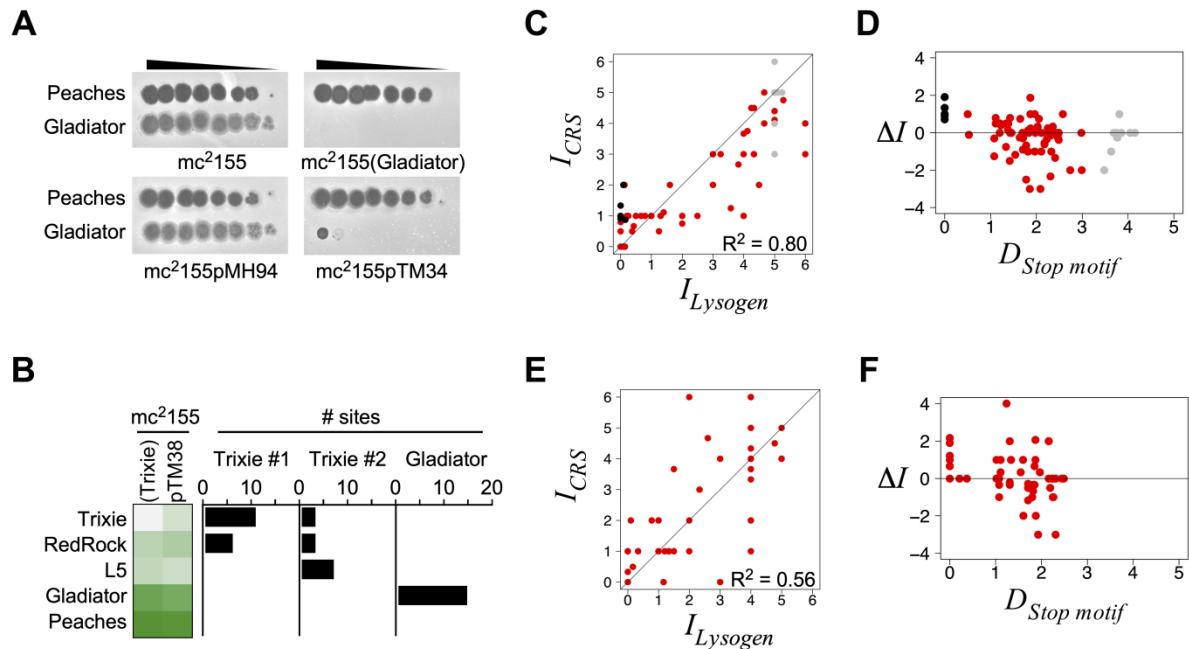


Figure 5-14. Mesoimmunity patterns are repressor-mediated.

(A) Representative immunity assays as in Figure 5-7, against mc²155, a lysogen (Gladiator), or CRSs (pMH94: empty vector; pTM34: Gladiator). (B) (left) Heatmap of infection phenotypes of Trixie, RedRock, L5, Gladiator, and Peaches against a Trixie lysogen and CRS (pTM38), as in Figure 5-10. (right) Horizontal histogram displaying the number of 13 bp stoperator sites present in each of the challenging phage genomes that match the stoperator sites in the indicated 30 bp EMSA substrates tested for Rep_{Trixie} binding affinity in Figure 5-6. (C) Scatter plot comparing superinfection scores of environmentally isolated phages against lysogens ($I_{Lysogen}$) and the corresponding CRS (I_{CRS}). The color scheme and R^2 values are as in Figure 5-12. The $y = x$ line is plotted for reference. Number of comparisons = 80. (D) Scatter plot comparing the change in infection scores (ΔI) between lysogens and the corresponding CRSs in panel C and stoperator motif distance ($D_{Stop\ motif}$). (E-F) Scatter plots as in panels C and D using lab-derived mutant phages. Number of comparisons = 54.

Divergence between repressor-mediated immunity may not always be a linear evolutionary path towards heteroimmunity. Et2Brutus is heteroimmune with DarthPhader (Subcluster A12), EagleEye (Subcluster A16), and Alma and Pioneer (Subcluster A9), but it is mesoimmune with DaVinci (Subcluster A6) and homoimmune with Gladiator (Subcluster A6)(Figure 5-15). However, the immunity patterns of an Et2Brutus lysogen and CRS are not congruent with the repressor nucleotide phylogeny. Either a distant ancestor at the root of the tree evolved heteroimmunity with Et2Brutus and subsequent evolution of the Subcluster A6 phages has re-subjected them to Et2Brutus immunity, or heteroimmunity along this evolutionary branch has independently evolved multiple times.

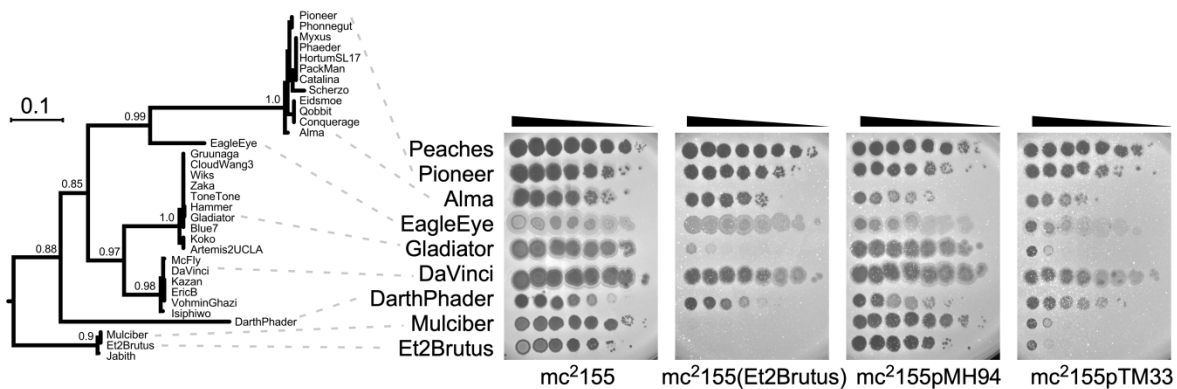


Figure 5-15. Evolutionary immunity transitions may not be linear.

Correlation of superinfection phenotypes involving a subset of L5 clade phages against an Et2Brutus lysogen and CRS (pTM33) with a phylogeny of immunity repressors constructed using maximum likelihood based on codon alignment. Peaches serves as a heterotypic control. Branch support values reflect aLRT.

Despite the strong correlation between lysogen and CRS immunity profiles, some phages produced discrepant infection phenotypes (Figure 5-14C). In general, infection of CRSs is equal to or weaker than infection of lysogens, and the most discrepant phenotypes occur at greater genetic distances (Figure 5-14D). This could be due to increased Rep expression in the CRS than in the lysogen resulting in stronger defense, similar to observations with CI on λ immunity (Bailone and Devoret, 1978). Alternatively, this could be due to genetic factors in the prophage that interfere with Rep's ability to inhibit lytic gene expression of the challenging phage.

5.3.5 Characterization of defense escape mutants

Virulent mutants that escape homotypic immunity have been characterized for several immunity systems, including λ (Bailone and Devoret, 1978; Lederberg and Lederberg, 1953), P22 (Bronson and Levine, 1971), P1 (Yarmolinsky, 2004), and Mu (Howe and Bade, 1975), but the relationship between homotypic and mesotypic virulence is not known. Virulent Cluster A phages have not been previously isolated (Donnelly-Wu et al., 1993), and no distinct plaques are observed during homotypic challenge of any of the prophages across the L5 clade. However, potential escape mutants are commonly observed in immunity assays involving superinfection of mesotypic CRSs and lysogens (Supplementary Table 5-1). Therefore, to determine how Cluster A phages can escape immunity from prophages within a mesoimmunity group, I isolated and characterized nine defense escape mutants (DEMs) that escape defense from six different lysogens or CRSs (Figure 5-16A, B). Dan Russell and Rebecca Garlena sequenced and assembled the DEM genomes.

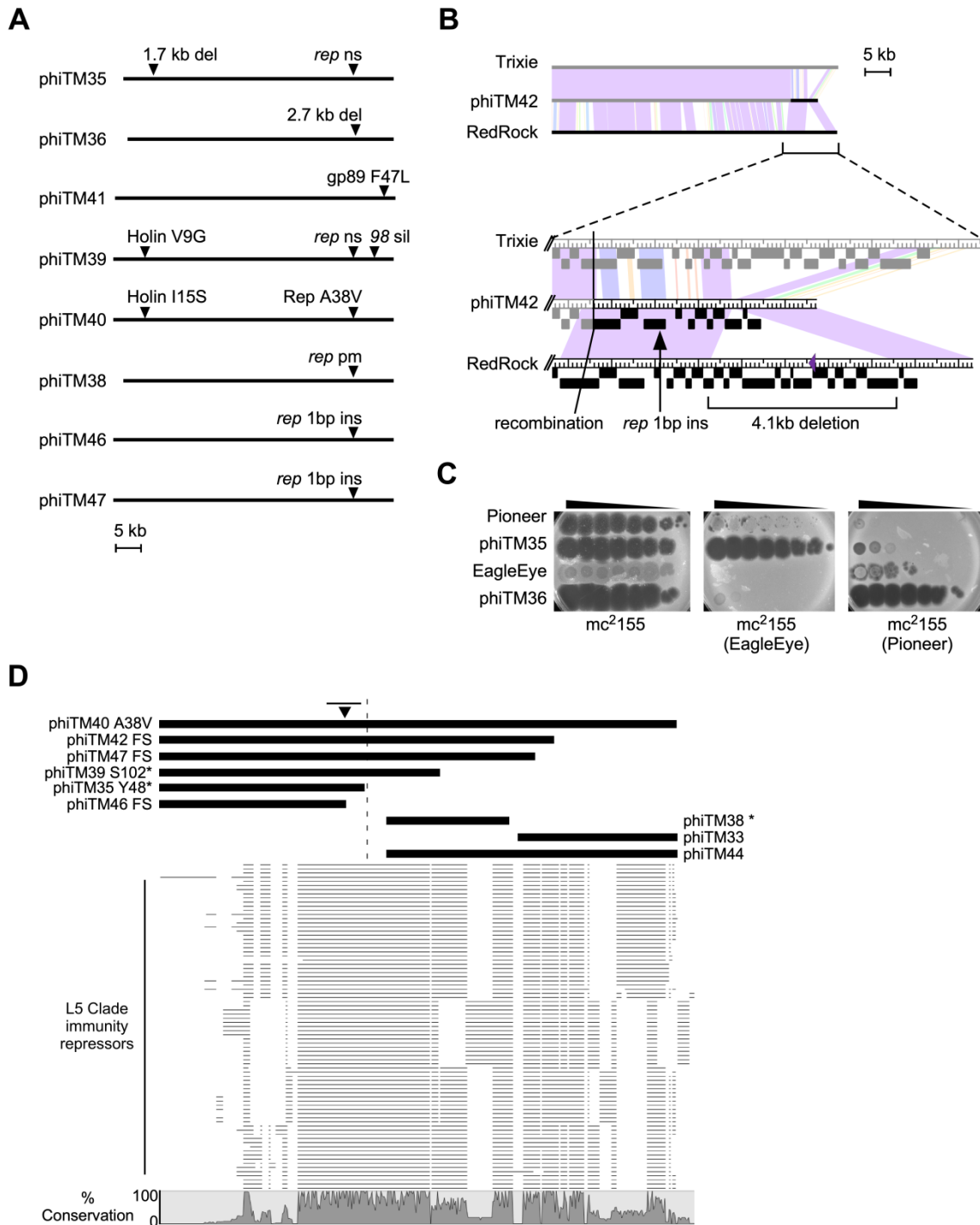


Figure 5-16. Summary of mutations present in defense escape mutants.

(A) Genome maps of DEMs that have escaped immunity from either a lysogen or CRS. Arrowheads indicate mutations. ns: nonsense; sil: silent; ins: insertion; del: deletion; pm: point mutation. **(B)** (top) Whole genome alignment of RedRock (black), Trixie (grey), and DEM phiTM42 (black and grey) using Phamerator. The color spectrum between genomes indicates sequence similarity (violet: significant similarity; white: no similarity).

(bottom) Enlarged view of the right genome termini, indicating the three mutations present in phiTM42 relative to Trixie and RedRock. **(C)** Immunity assay of Pioneer, phiTM35, EagleEye, and phiTM36 on mc²155, an EagleEye lysogen, and a Pioneer lysogen. **(D)** Codon alignment of repressors from L5 clade phages using PRANK. Nucleotide conservation indicated below. Dashed line separates predicted N-terminal and C-terminal domains. The position of the DNA-binding helix-turn-helix domain is indicated by thin black bar at top. Thick black bars above alignment indicate the repressor coding sequence present in several DEMs or isolated lytic mutants, highlighting gene truncations due to deletions, nonsense mutations (*), or frame shifts (FS). Black arrowhead indicates point mutation in phiTM40.

The DEMs contain a variety of mutations. An L5 mutant, phiTM41, escapes a Trixie lysogen after acquiring a single missense mutation in the first coding sequence downstream of *P_{left}*, gene 89 (Figures 5-9A, 5-16A). More substantial deletions are incurred by Pioneer and EagleEye mutants, phiTM35 and phiTM36, to escape each other's lysogen (Figure 5-16C). All but phiTM41 contain a mutation at the *rep* locus (Figure 5-16D). The most dramatic mutation is observed in phiTM42, isolated from a RedRock infection of a Trixie lysogen (Figure 5-16B). This DEM is a recombinant hybrid of the two phages, in which Trixie has lost the right most ~ 10 kb of its genome (including *P_{left}* and *rep*) and has acquired the analogous locus from RedRock. Over 4 kb of the RedRock fragment has been deleted and *rep* has acquired a 1 bp insertion. Similar to environmentally isolated temperate phages, these DEMs tend to exhibit decreased infection strength on CRSs compared to lysogens, which correlates with their genetic distance (Figure 5-14E, F).

I compared the infection strengths between each DEM and parent phage across a panel of lysogens to investigate how the mutations impact escape from other mesotypic prophages. DEMs nearly always exhibit equal or greater infection strengths than their parent phages on both lysogens and CRSs (Figure 5-17A). However, they are not able to escape all mesotypic immunity systems, and escape specificity is variable (Figure 5-17B). For example, the mutation

in phiTM41 does not impact infection of any lysogens other than Trixie (Figure 5-9A, see Figure 5-24B). phiTM41 is the only DEM with no mutation in *rep*, and although it forms slightly less turbid plaques than L5 it remains temperate (see Figure 5-23B). phiTM41 lysogens exhibit the same defense profile as an L5 lysogen, indicating that the missense mutation in gene 89 abolishes the asymmetric infection observed between Trixie and L5 without altering other infection or defense capabilities (Figure 5-9A, see Figure 5-24B). In contrast, the DEM phiTM42 is lytic and exhibits strong homotypic and mesotypic virulence: it escapes immunity from lysogens of both parent phages, Trixie and RedRock, and a Trixie CRS as well as every other lysogen tested (Figure 5-18A, B).

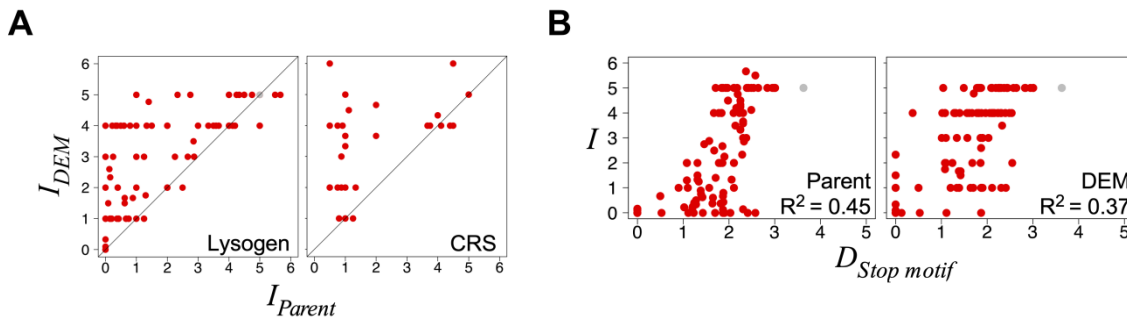


Figure 5-17. Defense escape mutants exhibit varying degrees of virulence.

(A) Scatter plots comparing infection scores of DEMs (I_{DEM}) and their parent phages (I_{Parent}) against (left, $n = 123$) lysogens or (right, $n = 30$) CRSs. The color scheme as in Figure 5-12. The $y = x$ line is plotted for reference. (B) Scatter plots comparing infection scores to stopoperator motif distances ($D_{Stop\ motif}$) for DEMs and their parent phages. The color scheme and R^2 values are as in Figure 5-12. Number of comparisons = 123.

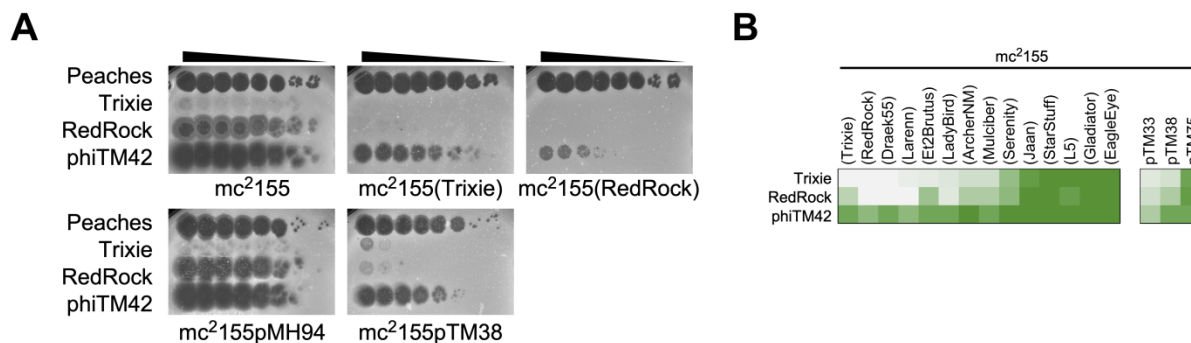


Figure 5-18. phiTM42 exhibits strong homotypic and mesotypic virulence.

(A) Representative immunity assays comparing infection phenotypes of Trixie, RedRock, and phiTM42 against mc²155, lysogens (Trixie and RedRock), and CRSs (pMH94: empty vector; pTM38: Trixie). (B) Heatmap of infection phenotypes as in Figure 5-10 comparing the infection profiles of phiTM42 with its parent phages against lysogens and CRSs (pTM33: Et2Brutus; pTM38: Trixie; pTM75: L5). Columns are ordered by increasing infection strength of the parent phages.

Similar types of escape mutations have variable impact on virulence specificity. phiTM39 and phiTM40 are Et2Brutus mutants that have escaped L5 and Trixie lysogens, respectively (Figure 5-16A). They both contain mutations in *rep* and *holin* genes, and both escape an L5 and Trixie CRS, respectively (Figure 5-19A, B). However, phiTM39 exhibits a broader virulence profile than phiTM40; although phiTM40 is now able to escape a Jaan lysogen, phiTM39 is now able to escape a Jaan lysogen, a Trixie lysogen, and a Trixie CRS. Additionally, unlike phiTM40, phiTM39 produces plaques on an Et2Brutus CRS, and exhibits a weak homotypic virulent phenotype, producing very tiny, barely detectable, plaques on an Et2Brutus lysogen.

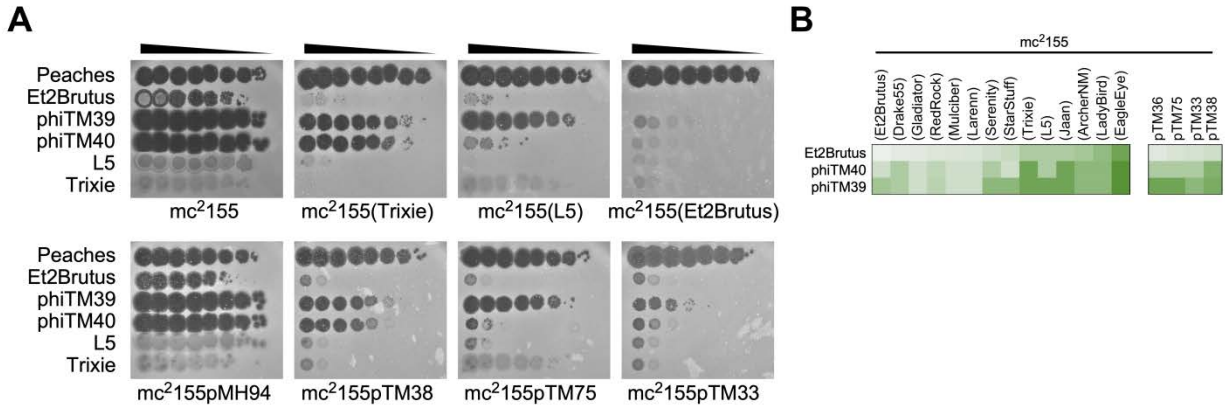


Figure 5-19. phiTM39 and phiTM40 exhibit different degrees of virulence.

(A) Representative immunity assays comparing infection phenotypes of Et2brutus, phiTM39, and phiTM40 against mc²155, lysogens (Trixie, L5, and Et2Brutus), and CRSs (pMH94: empty vector; pTM38: Trixie; pTM75: L5; pTM33: Et2Brutus). L5 and Trixie phages serve as negative controls for lysogens and CRSs. (B) Heatmap of infection phenotypes as in Figure 5-10 comparing the infection profiles of phiTM39 and phiTM40 with their parent phage, Et2Brutus, against lysogens and CRSs (pTM33: Et2Brutus; pTM36: StarStuff; pTM38: Trixie; pTM75: L5). Columns are ordered by increasing infection strength of the parent phage.

Similar effects are observed with derivatives of Gladiator and DaVinci, two closely-related Subcluster A6 phages (Figure 5-20A). They have very similar stoperator motifs and similar, but distinct, infection profiles (Figures 5-4C, 5-10). DaVinci and Gladiator mutants (phiTM46 and phiTM47, respectively) escape defense of a Gladiator CRS (mc²155pTM34) after acquiring a 1 bp insertion within *rep* (Figure 5-16A). The infection profile of phiTM46 against lysogens does not substantially differ from DaVinci (Figure 5-20B). In contrast, phiTM47's infection profile is now more similar to DaVinci and phiTM46 than to Gladiator, superinfecting Et2Brutus, Larenn, Mulciber, and Drake55 lysogens (Figure 5-20B). The infection profiles of phiTM46 and phiTM47 now are very similar to the infection profile of Jeffabunny (Subcluster A6), which is an obligately lytic mutant that has completely lost *rep* (Figure 5-20A, B).

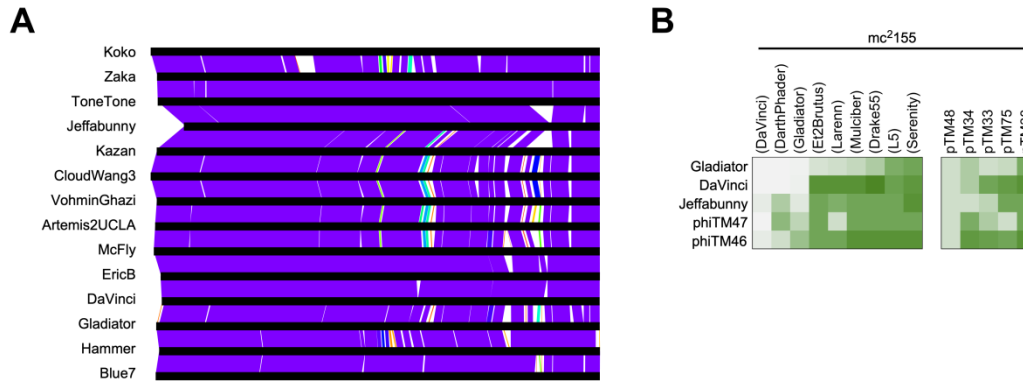


Figure 5-20. Comparison of phiTM46 and phiTM47 infection profiles.

(A) Phamerator whole genome alignment of Subcluster A6 phages. (B) Heatmap of infection phenotypes as in Figure 5-10 comparing the infection profiles of phiTM46 and phiTM47 with their parent phages against lysogens and CRSs (pTM33: Et2Brutus; pTM36: StarStuff; pTM38: Trixie; pTM48: DaVinci; pTM75: L5). Columns are ordered by increasing infection strength of the parent phage.

Furthermore, mutations conferring homotypic virulence do not necessarily confer mesotypic virulence. In contrast to phiTM42, which superinfects all lysogens tested, phiTM38 exhibits strong homotypic virulence but weak mesotypic virulence. This DEM is a derivative of phiTM44, which itself is >99% identical to the lytic phage D29 (Table 5-1). A single point mutation in phiTM38's *rep* locus confers escape from an Et2Brutus lysogen (Figure 5-16A). As a derivative of D29, phiTM38 is closely related to L5 (Ford et al., 1998), and it exhibits 98% sequence similarity to StarStuff, its closest temperate relative in the actinobacteriophage database (Figure 5-2). Surprisingly, phiTM38 can escape immunity from L5 and StarStuff lysogens and CRSs (Figure 5-21A, B). Despite this virulence, phiTM38 remains unable to superinfect lysogens of more distantly-related phages, such as Larenn (Figure 5-21A).

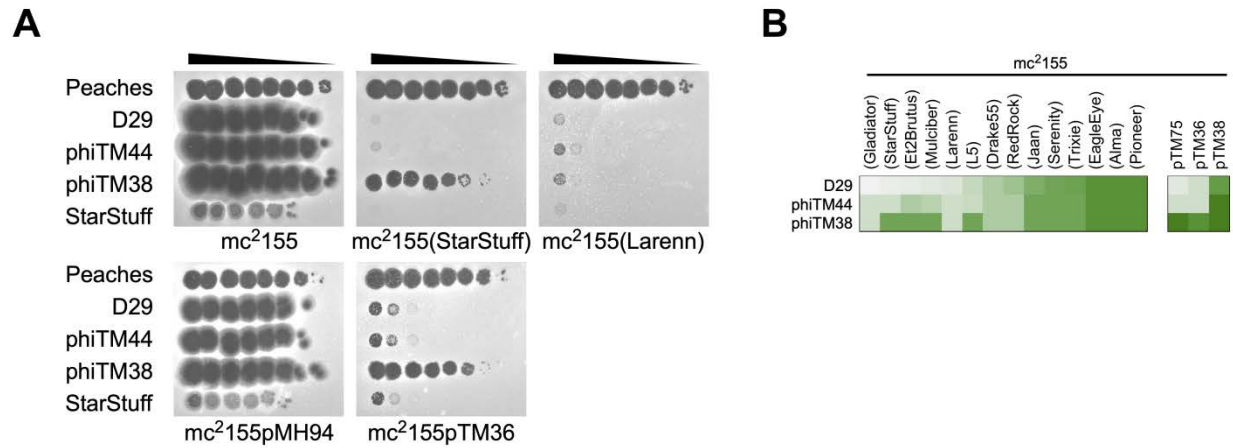


Figure 5-21. phiTM38 exhibits limited virulence.

(A) Representative immunity assays involving D29, phiTM44, and phiTM38 against mc^2155 , lysogens (StarStuff and Larenn), and CRSs (pMH94: empty vector; pTM36: StarStuff). StarStuff serves as a negative control for lysogens and CRSs. (B) Heatmap of infection phenotypes as in Figure 5-10 comparing the infection profiles of phiTM38 with its parent phages against lysogens and CRSs (pTM36: StarStuff; pTM38: Trixie; pTM75: L5). Columns are ordered by increasing infection strength of the parent phage.

5.3.6 An engineered L5 mutant exhibits acute homotypic virulence

The immunity repressors of lambdoid phages are similarly structured, containing a helix-turn-helix domain near the N-terminus responsible for DNA binding and a C-terminal domain responsible for dimerization that impacts sequence specificity (Donner et al., 1998; Sauer et al., 1982; Valenzuela and Ptashne, 1989). The Cluster A Rep harbors an N-terminal helix-turn-helix DNA-binding domain as well (Donnelly-Wu et al., 1993; Pope et al., 2011b) and may also contain a distinct C-terminal domain (Ganguly et al., 2007). Nucleotide sequence alignment of the immunity repressors from the L5 clade suggest these two regions are under markedly different evolutionary pressures (Figure 5-16D). All repressors exhibit a region of high sequence similarity that extends across the entire N-terminal domain and into the C-terminal domain. In contrast, alignment of the C-terminal domain is much poorer and exhibits more sequence variation. The pairwise amino acid genetic distances between the C-terminal domains are larger than the distances between the N-terminal domains (Figure 5-22A). Surprisingly, asymmetric and incomplete immunity phenotypes correlate with genetic diversity of the repressor C-terminus, instead of with the N-terminus or the specific helix-turn-helix domain (Figure 5-22B). For instance, the N-terminus of Rep_{StarStuff} is 97% similar to Rep_{Gladiator} and Rep_{DaVinci}, including an identical helix-turn-helix domain, but their stoperator motifs are distinct, and they exhibit asymmetric superinfection phenotypes on lysogens and CRSs (Figure 5-22C, D). Therefore, the C-terminus of the Cluster A Rep may play an important role in immunity, as in the lambdoid immunity systems.

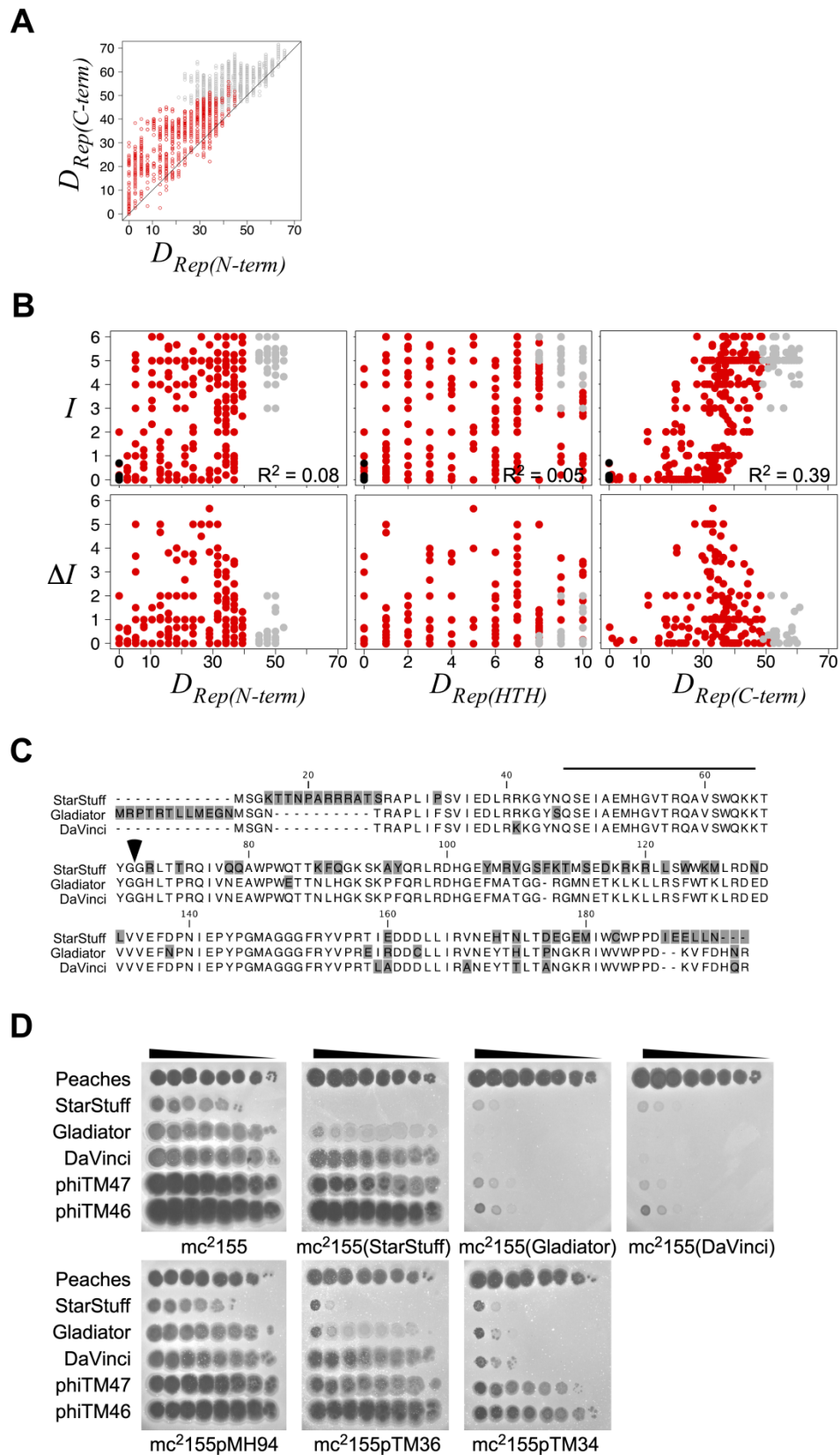


Figure 5-22. Evolution of immunity repressor domains.

(A) Scatter plot comparing repressor N-terminus ($D_{Rep(N-term)}$) and C-terminus ($D_{Rep(C-term)}$) genetic distances for 87 L5 clade phages. **(B)** Scatter plots comparing (top, $n = 423$) the averaged infection score (I) or (bottom, $n = 185$) the infection score difference (ΔI) between reciprocal assays to the genetic distance of the repressor N-terminal domain ($D_{Rep(N-term)}$), C-terminal domain ($D_{Rep(C-term)}$), or the hamming distance of the predicted helix-turn-helix domain ($D_{Rep(HTH)}$). For top row plots, the R^2 value from a linear regression is indicated. **(C)** Alignment of StarStuff, Gladiator, and DaVinci Rep homologs, with the helix-turn-helix domain indicated by a black bar, the N-terminal and C-terminal regions delineated by an arrow, and amino acid variants shaded in grey. **(D)** Immunity assays involving StarStuff, Gladiator, DaVinci, phiTM47, and phiTM46 against mc²155, lysogens (StarStuff, Gladiator, and DaVinci), and CRSs (pMH94: empty vector; pTM36: StarStuff; pTM34: Gladiator). Peaches serves as a heterotypic control.

To study the impact of the repressor C-terminus in immunity, I characterized L5 derivatives that I engineered to contain an immunity repressor with either a C-terminal FLAG (phiTM6) or HA (phiTM1) tag (Figure 5-23A). Compared to L5, phiTM1 produces slightly darker plaques while phiTM6 produces more turbid plaques (Figure 5-23B). Lysogens of both engineered mutants can be generated, and they exhibit lower levels of spontaneous phage release than L5 or phiTM41 (data not shown). The repressor expressed from a phiTM6 prophage contains the translated FLAG tag, as it can be detected by Western blot (Figure 5-23C). Infection phenotypes of both engineered mutants do not substantially differ from L5 or phiTM41 infection phenotypes when tested against a panel of lysogens and CRSs (Figures 5-24A, B, 5-25A). Similarly, the infection phenotypes of diverse Cluster A phages against the mutant L5 lysogens do not substantially differ compared to L5 or phiTM41 lysogens, although there may be a few phages that are slightly impacted (Figures 5-24B, 5-25B). Thus, although the C-terminal tags slightly impact phage growth, they do not appear to have a substantial impact on infection or immunity.

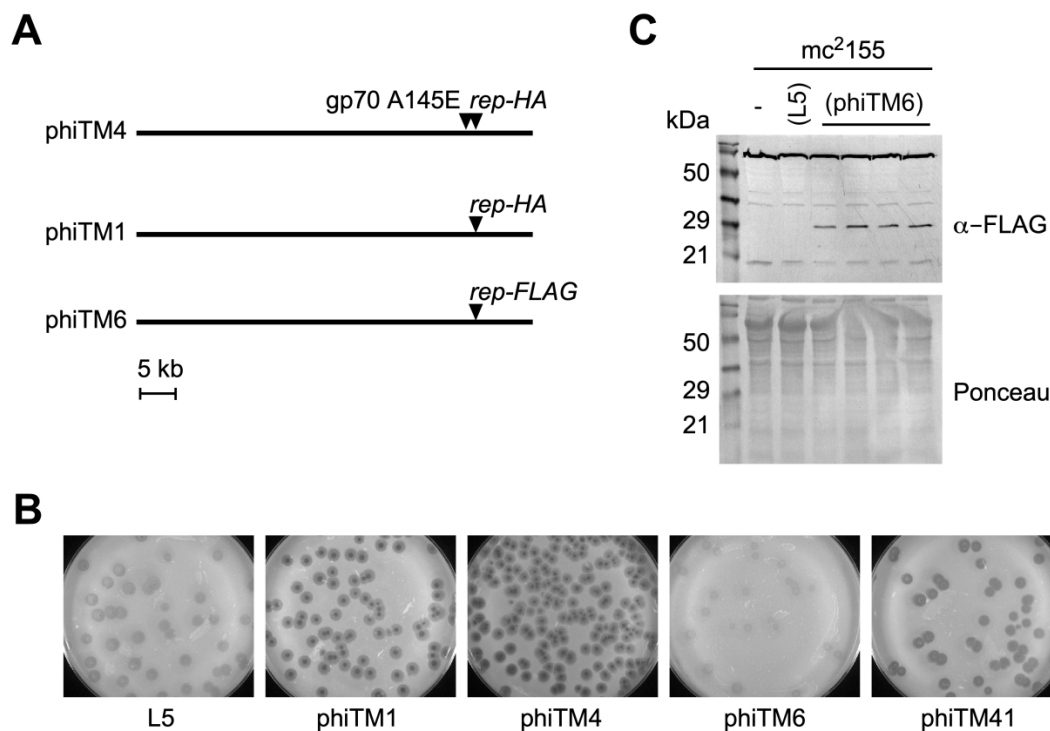


Figure 5-23. Characterization of L5 engineered mutants and defense escape mutants.

(A) Genome maps of L5 and derivative mutants indicating the engineered mutations present in phiTM6 and phiTM1, and the unintentional mutation acquired in phiTM4. (B) Comparison of plaque morphologies from wild type L5 and several L5 derivatives. (C) (top) Western blot detecting Rep-FLAG (expected size 22 kDa) in *mc*²¹⁵⁵, an L5 lysogen, and several independent replicate FLAG-tagged L5 lysogen cultures. (bottom) Ponceau stain of identical gel measuring total protein content per lane.

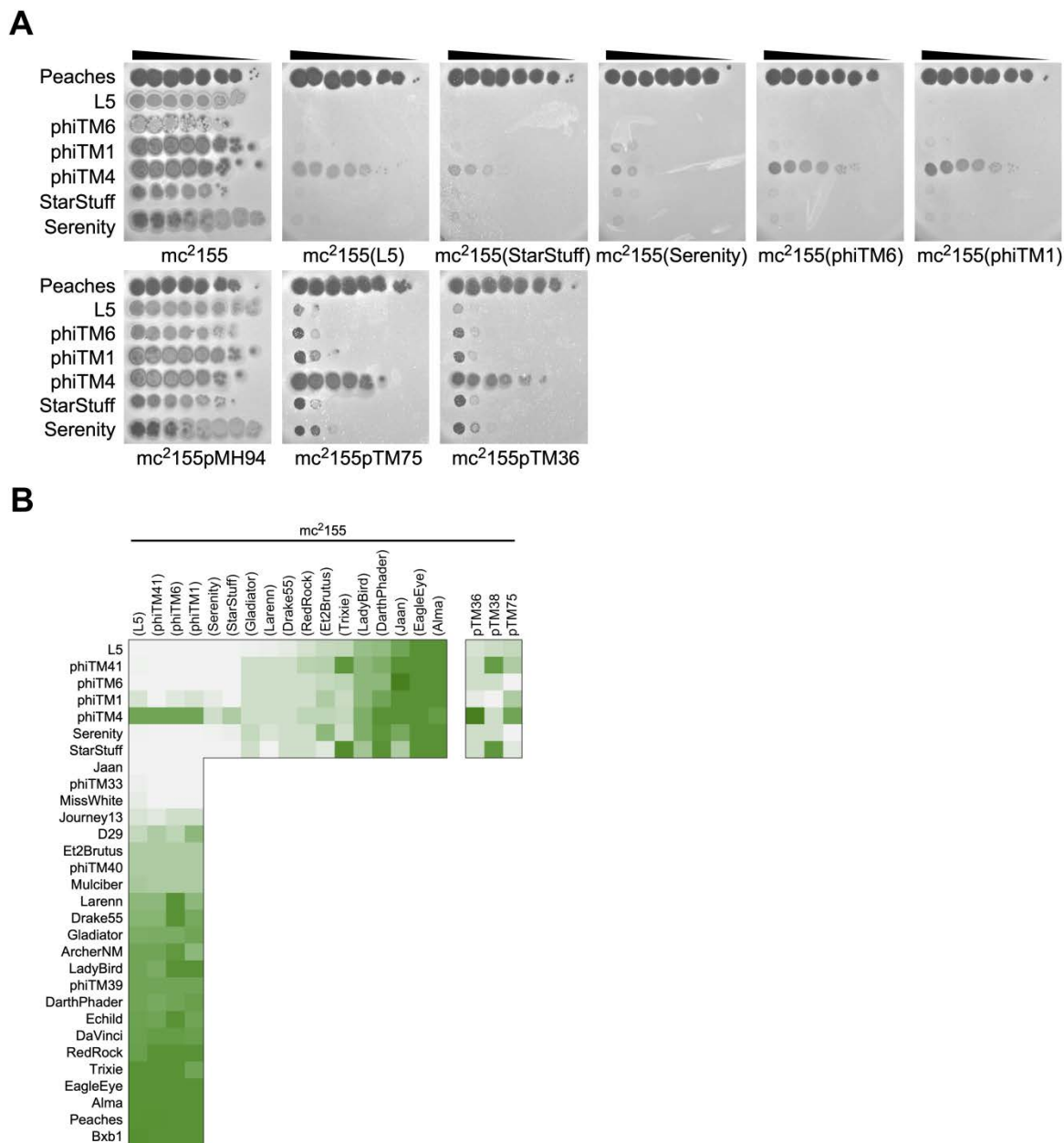


Figure 5-24. Superinfection and immunity profiles of L5 engineered mutants.

(A) Representative immunity assays comparing infection phenotypes of L5 and several L5 derivatives (phiTM6, phiTM1, and phiTM4) against mc²155, lysogens (L5, StarStuff, Serenity, phiTM6, and phiTM1), and CRSs (pMH94: empty vector; pTM75: L5; pTM36: StarStuff). Peaches serves as a heterotypic control, and Serenity and StarStuff serve as negative controls for lysogens and CRSs. **(B)** Heatmap of infection phenotypes as in Figure 5-10 comparing infection profiles of phages against L5, phiTM41, phiTM1, and phiTM6 lysogens, and the infection profiles of L5, phiTM41, phiTM1, phiTM4, phiTM6, Serenity, and StarStuff against several lysogens and CRSs (pTM75: L5; pTM36: StarStuff; pTM38: Trixie; pTM34: Gladiator). Rows are ordered by increasing infection strength on an L5 lysogen, and columns are ordered by increasing L5 infection strength.

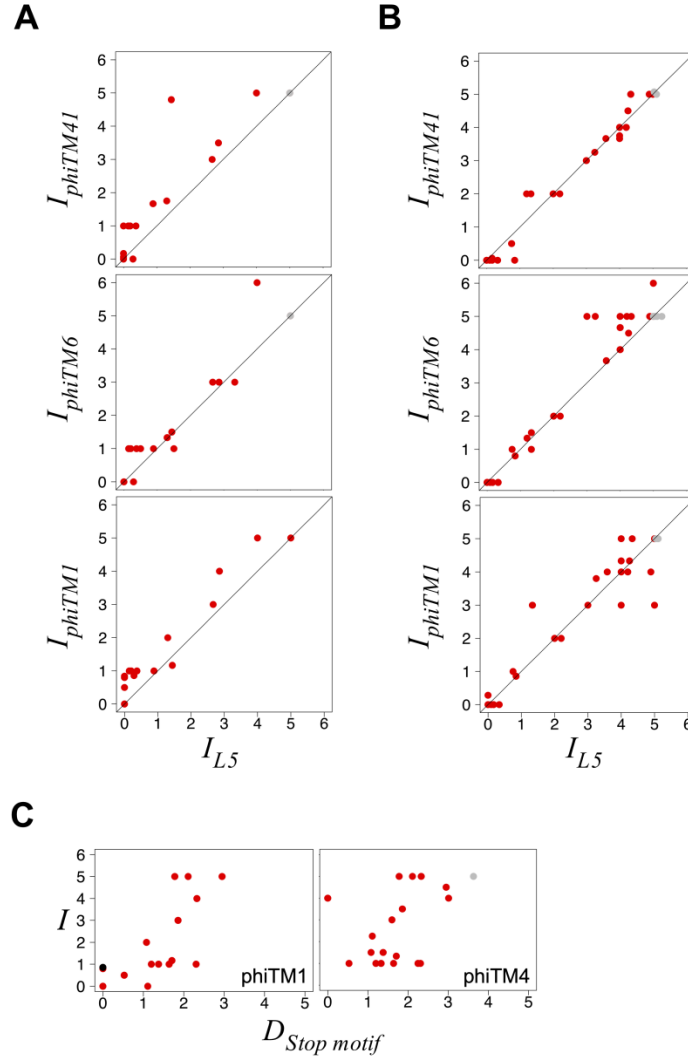


Figure 5-25. phiTM4 exhibits limited virulence.

(A) Scatter plots comparing infection phenotypes of L5 phage (I_{L5}) and L5 mutant phages against lysogens or CRSs. The $y = x$ line is plotted for reference. Number of comparisons = 25. (B) Scatter plots comparing infection phenotypes of phages against an L5 lysogen (I_{L5}) or L5 mutant lysogens. Number of comparisons = 44. (C) Scatter plots comparing infection scores to the stopoperator motif distance ($D_{Stop\ motif}$) of phiTM1 and phiTM4 infections against lysogens and CRSs.

During purification of a phiTM1 lysogen, a homotypic virulent mutant derivative, phiTM4, was inadvertently isolated. Dan Russell and Rebecca Garlena sequenced and assembled this genome. phiTM4 has acquired a point mutation within gene 70 immediately downstream of rep (Figure 5-23A). The impact of this mutation is not certain. The point mutation changes the

last amino acid in gp70. The point mutation also occurs within a stoperator site 1 bp upstream of the start of the highly conserved gene 69. Neither gene has a known function, although 69 exhibits sequence similarity to Cas4-like genes. phiTM4 superinfects lysogens of all L5 derivatives, including phiTM1, a StarStuff lysogen, and L5 and StarStuff CRSs (Figure 5-24A, B). However, it remains unable to superinfect other mesotypic lysogens, including its closest relative in the database, Serenity (Figures 5-24A, 5-25C). The single point mutation confers the most acute homotypic virulence observed, in which phiTM4 escapes homotypic immunity from a nearly identical prophage but remains subject to many other mesotypic immunity systems.

5.3.7 Evolution of stoperators

The genetic diversity present in L5 clade phages also enables deeper investigation into the evolution of stoperators at P_{left} that was not previously possible. In contrast to previously described genetic relationships between Cluster A phages such as L5, D29, and Bxb1, the subclade of D29 relatives contain genome sequences sufficiently similar to analyze by whole genome alignment and phylogenetic reconstruction (Figure 5-2A, B). Alignment of these phages provides insight into how L5 clade phage genomes evolve. Among the D29 related phages, there are ~ 500-1000 single nucleotide point mutations and 25 insertions and deletions [Tables 5-3, 5-4, originally published in (Dedrick et al., 2017b)]. However, the mutations are not randomly distributed throughout the genome (Figure 5-26A). Particular loci appear to accumulate more point mutations than others, such as near the tape measure protein, Rep, and P_{left} . The majority of insertions and deletions, incurred along the Kerberos branch, occur at the right end of the genome (Figure 5-26A, B). However, there are few mutations within stoperators [Table 5-5, originally published in (Dedrick et al., 2017b)].

Table 5-3. D29 sub-clade single nucleotide polymorphisms.

	Kerberos	Pomar16	StarStuff
D29	950	737	744
Kerberos		587	685
Pomar16			546

Table 5-4. D29 sub-clade insertions and deletions.

Insertion/ Deletion #	Insertion/ Deletion	Phylogenetic Branch	Size (nt)	StarStuff	D29	Pomar16	Kerberos	Note
1	insertion; deletion	Pomar16	47; 5	4,399		4,399		
2	deletion	Pomar16	1	4,498		4,540		
3	insertion	Kerberos	3	15,679			15,682	
4	insertion	StarStuff	3	25,302				
5	insertion	D29	1	32,068	32,074			
6	deletion	Pomar16	1	45,317		45,355		
7	deletion	D29	3,668	45,676	45,583			repressor locus deletion
8	insertion	StarStuff	2	45,903				inside D29 deletion
9	insertion; deletion	Kerberos	3; 33	46,136			46,131	inside D29 deletion
10	insertion	Pomar16	3	46,670		46,705		inside D29 deletion
11	insertion	Kerberos	3	51,288			51,253	
12	insertion	Kerberos	2	51,407			51,375	
13	insertion	Kerberos	2	51,413			51,383	
14	insertion	Kerberos	1	51,418			51,390	
15	deletion	Kerberos	11	51,432			51,405	
16	insertion	Kerberos	7	51,580			51,542	
17	insertion	Kerberos	5	51,588			51,557	
18	deletion	Kerberos	1	51,604			51,578	
19	deletion	Kerberos	1	51,619			51,592	
20	insertion	Kerberos	1	51,629			51,601	
21	insertion	ancestor of Kerberos and Pomar16	1	52,002		52,040	51,975	
22	deletion	StarStuff; Kerberos	11	52,389			52,361	
23	deletion	Kerberos	6	52,400			52,372	
24	insertion	Kerberos	2	52,539			52,505	
25	insertion	D29	1	52,680	49,030			

Note: genomic coordinates indicate the aligned nucleotide position adjacent to each event and are provided for the specific genome(s) in which the event occurred, as well as in StarStuff for reference.

Table 5-5. D29 sub-clade stopoperator sites.

D29 Site #	Strand	D29		StarStuff		Pomar16		Kerberos		Notes
		Left	Right	Left	Right	Left	Right	Left	Right	
1	bottom	48,536	48,548	52,197	52,209	52,234	52,246	52,169	52,181	operator, based on alignment to L5
2	bottom	48,423	48,435	52,084	52,096	52,121	52,133	52,056	52,068	
4	bottom	48,595	48,607	52,256	52,268	52,293	52,305	52,228	52,240	
5	bottom	48,296	48,308	51,957	51,969	51,995	52,007	51,930	51,942	
6	bottom	47,954	47,966	51,615	51,627	51,653	51,665	51,587	51,599	Kerberos differs from the others
7	bottom	47,613	47,625	51,274	51,286	51,312	51,324	51,239	51,251	
10	bottom	47,241	47,253	50,902	50,914	50,940	50,952	50,867	50,879	
12	bottom	44,042	44,054	44,035	44,047	44,073	44,085	44,032	44,044	
13	bottom	43,744	43,756	43,737	43,749	43,775	43,787	43,734	43,746	
14	bottom	41,878	41,890	41,871	41,883	41,909	41,921	41,868	41,880	
15	bottom	41,326	41,338	41,319	41,331	41,357	41,369	41,316	41,328	
16	bottom	39,116	39,128	39,109	39,121	39,147	39,159	39,106	39,118	
17	bottom	36,779	36,791	36,772	36,784	36,810	36,822	36,769	36,781	
18	bottom	32,881	32,893	32,874	32,886	32,912	32,924	32,871	32,883	
19	bottom	29,317	29,329	29,311	29,323	29,349	29,361	29,308	29,320	
20	top	19,026	19,038	19,017	19,029	19,058	19,070	19,017	19,029	
21	top	15,597	15,609	15,588	15,600	15,629	15,641	15,588	15,600	
22	top	13,059	13,071	13,050	13,062	13,091	13,103	13,050	13,062	
23	top	4,680	4,692	4,671	4,683	4,712	4,724	4,671	4,683	
24	bottom	199	211	190	202	190	202	190	202	
31	bottom	48,759	48,771	52,409	52,421	52,457	52,469	52,375	52,387	
32	bottom	47,507	47,519	51,168	51,180	51,206	51,218	51,133	51,145	
33	bottom	40,281	40,293	40,274	40,286	40,312	40,324	40,271	40,283	
34	bottom	46,294	46,306	49,955	49,967	49,993	50,005	49,920	49,932	StarStuff differs from the others
N/A	bottom	N/A	N/A	49,176	49,188	49,214	49,226	49,141	49,153	site is within D29 deletion locus

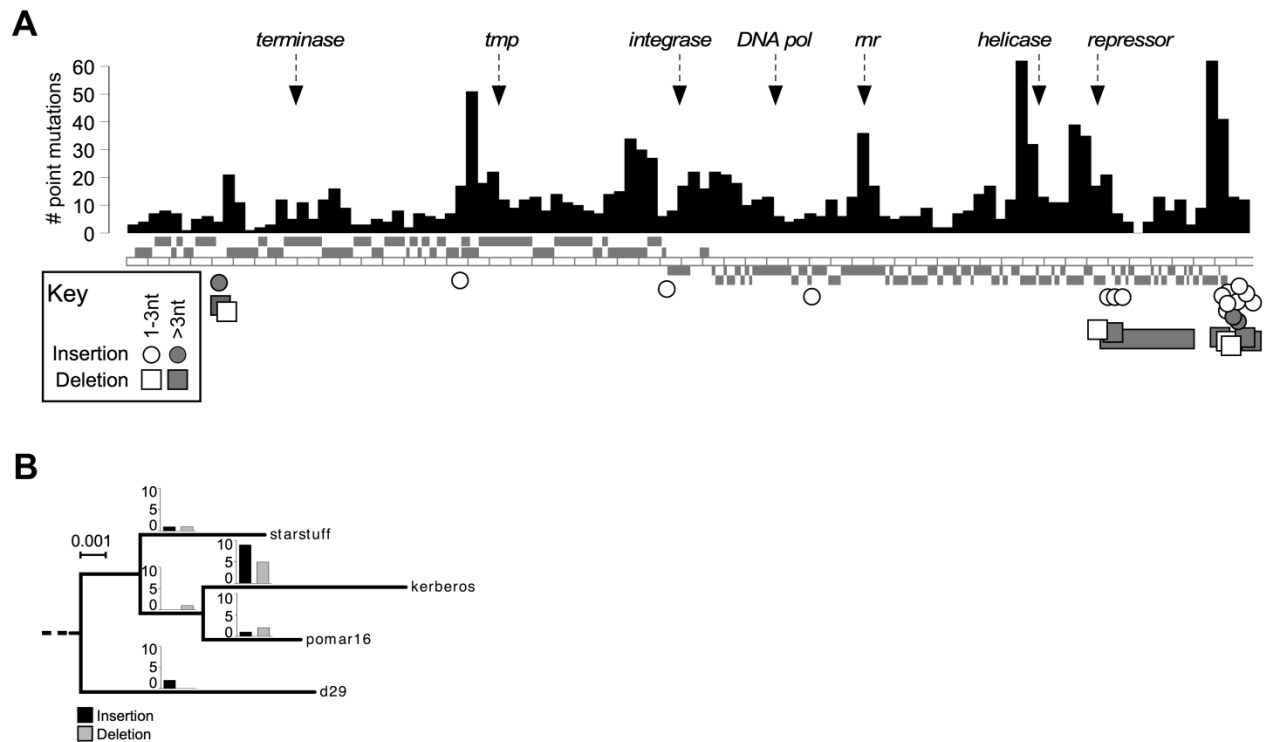


Figure 5-26. Genome evolution of D29 and its relatives.

(A) Diagram reflecting sequence mutations across the D29 sub-clade, mapped onto the StarStuff genome. Histogram reflects the number of SNPs identified. The individual insertion (circle) and deletion (square) events predicted from Count are labeled below and shaded by size. Several highly-conserved genes are labeled as in Figure 5-2 for reference. (B) Alignment gaps were mapped to the phylogenetic branches in the D29 sub-clade (enlarged from Figure 5-2B) using Count to predict whether gaps were due to insertion or deletion events. Bar charts reflect the total number of predicted events per branch. Figure adapted from (Dedrick et al., 2017b).

Greater sequence diversity of stoperators and operators at P_{left} is observed among Subcluster A6 phages though. Similar to D29 and its relatives, these phages exhibit significant sequence similarity across their genomes (Figure 5-20A). For example, Gladiator and DaVinci exhibit 93% sequence identity across 94% of their genomes. Alignment of their P_{left} loci indicates that of the five repressor binding sites, all but one have accumulated at least one point mutation in at least one genome (Figure 5-27A, B). Among the five binding sites specifically in Gladiator and DaVinci, three contain at least one point mutation, indicating they may play a role in the differences in superinfection immunity phenotypes (Figure 5-10).

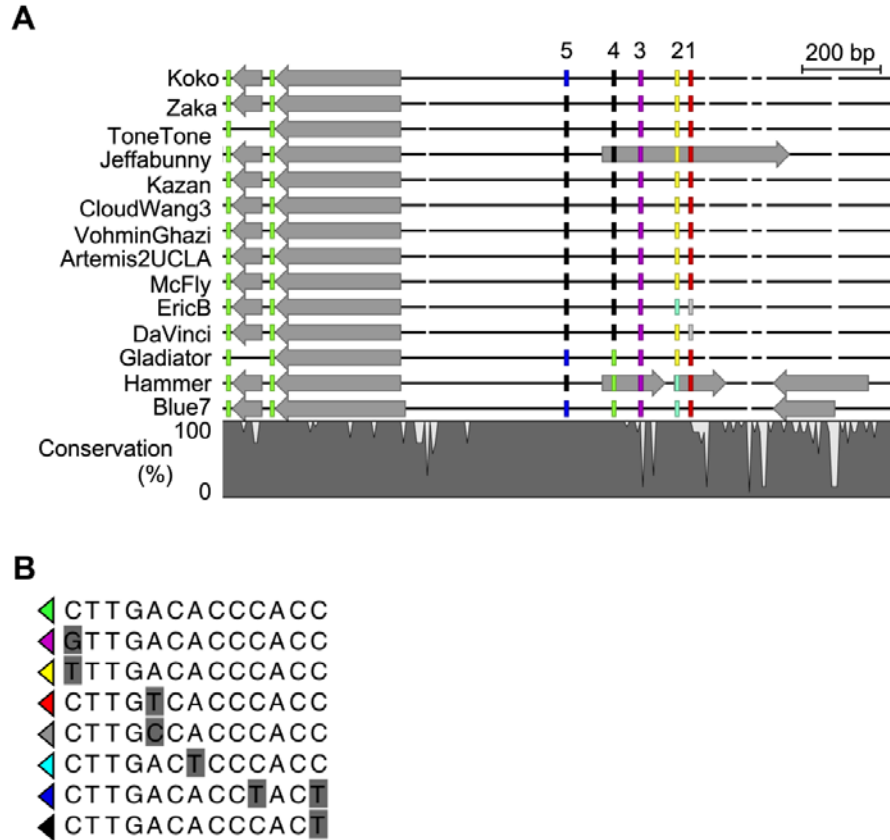


Figure 5-27. P_{left} stoperator conservation among Subcluster A6 phages.

(A) Enlarged view of the ~ 1.5 kb P_{left} locus from Figure 5-20A which has been aligned with MUSCLE. Genes are indicated by grey boxes. Stoperators are indicated by arrowheads, colored by their 13 bp sequence and numbered based on their order from the right genome termini. **(B)** Alignment of the eight unique stoperator sequences present in panel A, with corresponding color designations. Variant nucleotides are highlighted in grey.

Sequence diversity at the P_{left} locus of Subcluster A9 phages suggests that there may not be one particular promoter and operator. Similar to Subcluster A6 phages and the Subcluster A2 D29 sub-clade phages, Subcluster A9 phages exhibit significant sequence similarity across their genomes (Figure 5-28A). Alignment of the P_{left} locus from these genomes indicates that of the five predicted binding sites, Sites 4 and 5 have accumulated point mutations in at least one genome, and Site 2 has become deleted in phages Phonnegut and Pioneer (Figure 5-28B, C). This deletion may impact P_{left} expression (Figure 5-28D). In an Alma lysogen, expression is detected immediately downstream of Site 1. In a Pioneer lysogen, strong expression is detected immediately downstream of Site 3, and only very weak expression is barely detectable downstream of Site 1. Expression from P_{left} in an L5 lysogen occurs immediately downstream of the empirically determined operator (Figure 5-5C)(Brown et al., 1997). By comparison, Site 1 represents the operator in Alma, but Site 3 represents the operator in Pioneer (Figure 5-28D). This suggests that there are multiple promoters in the P_{left} locus of Subcluster A9 phages from which transcription can initiate, and the cluster of binding sites may enable the immunity repressor to block transcription regardless of where transcription begins.

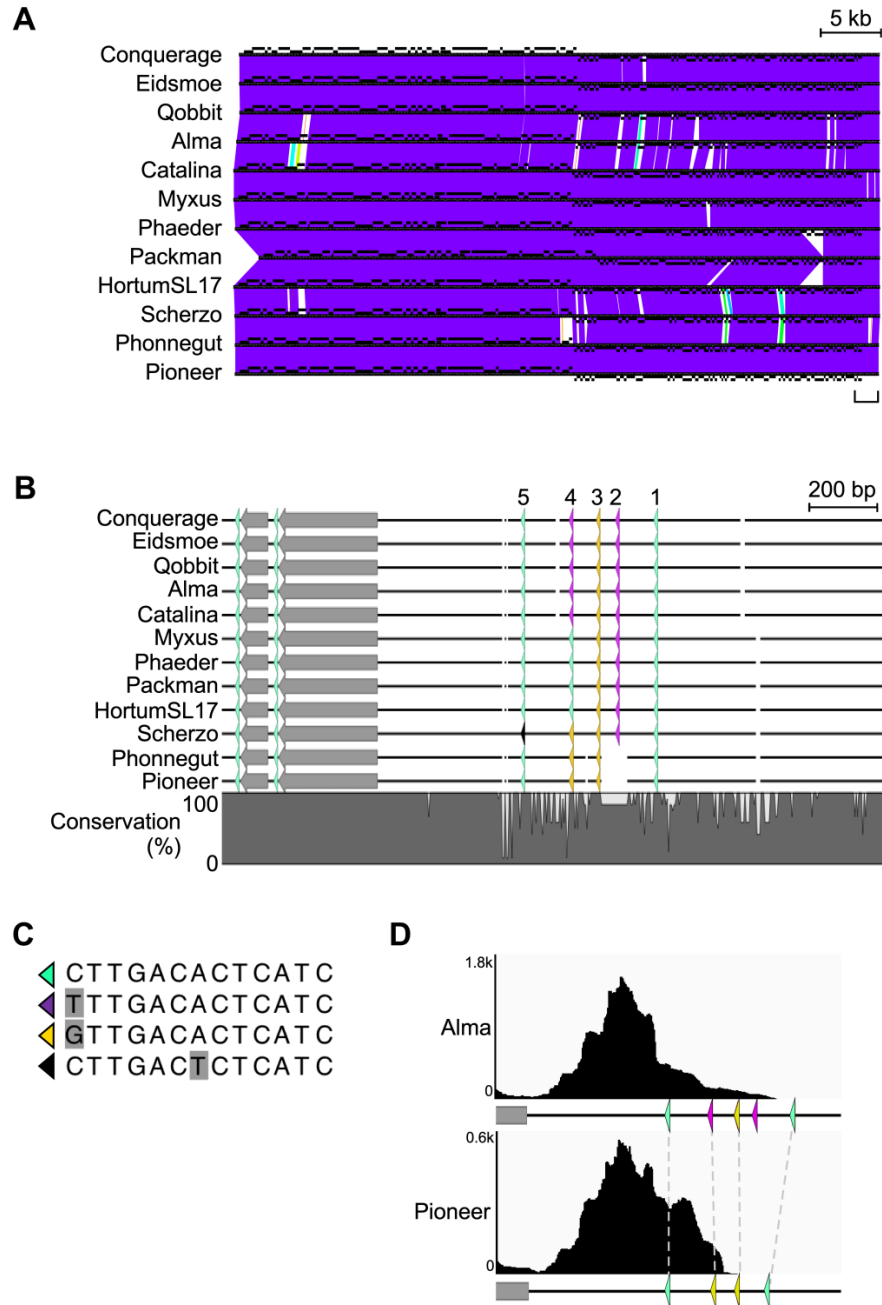


Figure 5-28. P_{left} expression and stoperator conservation among Subcluster A9 phages.

(A) Phamerator alignment of Subcluster A9 phages. (B) Enlarged view of the ~1.5 kb P_{left} locus from panel A and aligned with MUSCLE, with genes (grey boxes) and stoperators (arrowheads) indicated. Stoperators are colored by their 13 bp sequence and ordered from the right genome termini. (C) Alignment of the four unique stoperator sequences present in panel B, with corresponding colors, and with variant nucleotides (grey) indicated. (D) Enlarged view of the ~400 bp P_{left} locus from Alma and Pioneer manually aligned by Stoperator 5 with RNAseq profiles as in Figure 5-5. Dashed lines connect homologous stoperator from panel C.

5.3.8 Expression from P_{left} is toxic

Strong expression is observed downstream of the P_{left} locus during lysogeny and at all stages of infection for all phages tested (Figure 5-5C). Experiments performed by Rebekah Dedrick revealed that plasmids carrying a copy of the P_{left} locus and this highly transcribed region were unable to be transformed into *M. smegmatis*, suggesting that expression from P_{left} is toxic to the host (data not shown). I investigated this locus further by creating a series of plasmid constructs in which P_{left} is regulated by a thermo-inducible repressor (see Materials and Methods). I created a plasmid (pTM29 and pTM31) that carries a copy of the temperature sensitive L5 repressor allele (Donnelly-Wu et al., 1993). I then cloned a ~ 1.4 kb segment of the L5 P_{left} locus into the pTM29 vector (Figure 5-29). The cloned segment spans the promoter, the empirically determined operator, the cluster of repressor binding sites, the highly expressed region, gene 89 (which confers L5 escape from Trixie immunity), and ~ 400 bp downstream of gene 89. These constructs appear to be slightly toxic to *E. coli*, since all plasmid constructs recovered from *E. coli* transformations contained mutations in the P_{left} locus, ranging from single nucleotide deletions or point mutations (pTM12, pTM14, pTM8) to larger deletions (pTM9, pTM10, pTM11). In contrast to the P_{left}-only construct (tested by Rebekah Dedrick), the repressor-controlled P_{left} construct could successfully be transformed into *M. smegmatis*. Growth assays using this series of plasmids indicate that all recombinant strains grow well at 30°C, in which the repressor is stably expressed (Figure 5-29). However, when grown at 44°C, in which the L5 repressor is no longer stable, all recombinant strains carrying a repressor-P_{left} construct exhibit poorer growth than strains carrying the repressor-only construct (pTM29 or pTM31). These results suggest that expression of one or more genetic elements at the P_{left} locus are highly

toxic to *M. smegmatis*, and the repressor is required to regulate expression of this locus during lysogeny.

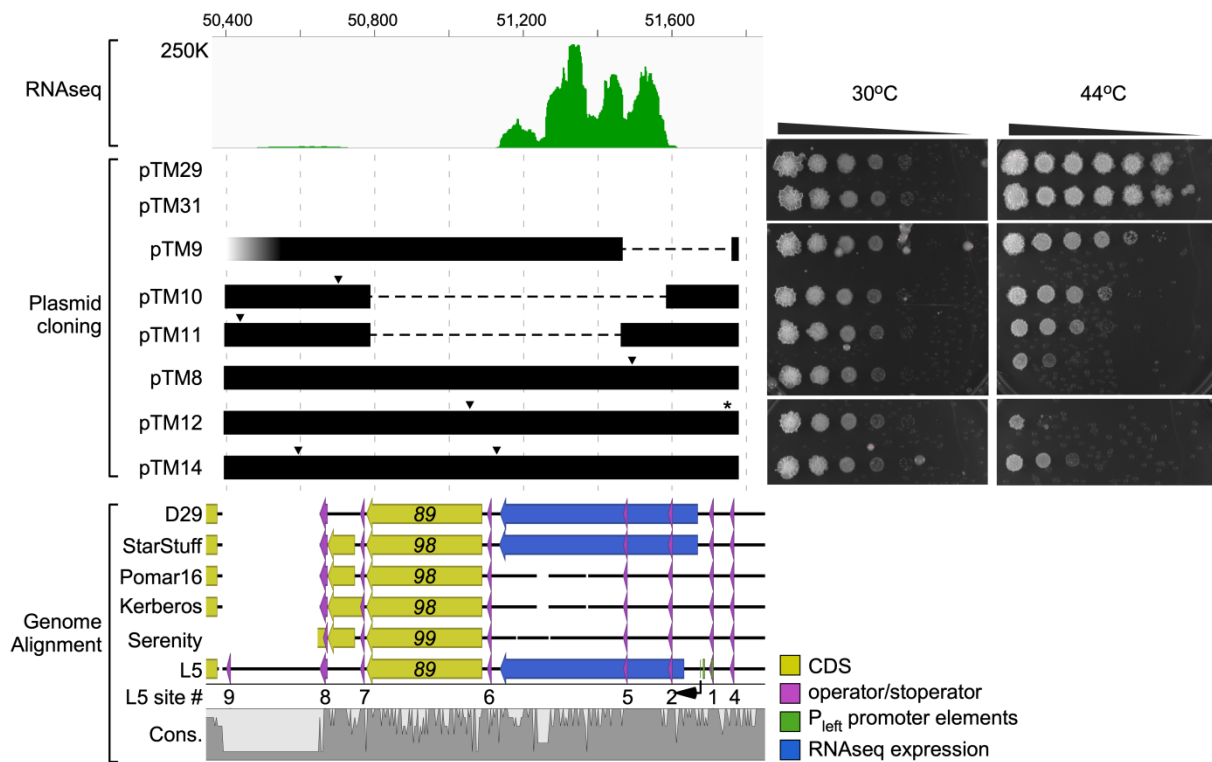


Figure 5-29. Toxicity of the highly expressed transcript from P_{left} .

(left, bottom) Alignment of the P_{left} locus from L5, D29, and D29 relatives indicating positions of stoperators, operators, P_{left} promoter elements, approximate RNAseq expression tracks (performed by Rebekah Dedrick, Dan Russell, and myself), and annotated coding sequences. Black lines indicate nucleotide sequence with alignment gaps, and the conservation track indicates levels of nucleotide identity across the alignment. **(left, middle)** Diagram of cloned L5 P_{left} sequence (black box) present in several plasmids and manually aligned to genome map below. Horizontal dashed lines indicate large deleted segments identified by sequencing. Triangles indicate single nucleotide mutations. The star indicates a single nucleotide deletion. The faded region in pTM9 reflects incomplete sequencing verification. Constructs also contain the cloned L5 integration and temperature-sensitive repressor allele loci, which are not shown. **(left, above)** RNAseq bottom strand expression profile of L5 during lysogeny manually aligned to genome map below. **(right)** Ten-fold serial dilutions of *M. smegmatis* cultures carrying plasmids indicated at left grown at 30°C and 44°C for four days with kanamycin selection. Figure adapted from (Dedrick et al., 2017b).

5.3.9 An extended cloned repressor locus strengthens immunity

Initial characterizations of the L5 immunity repressor showed that an *M. smegmatis* recombinant strain carrying a copy of the L5 repressor gene and upstream promoter in a ~ 1.3-1.5 kb segment in either an integrating or extrachromosomal vector was necessary and sufficient for conferring immunity against L5 (Donnelly-Wu et al., 1993). This cloned segment did not contain any other complete coding sequences from this locus, and immunity from the CRS did not appear as strong as from the cognate lysogen. Similar results are observed with the other L5 clade repressors I have cloned, such as Gladiator *rep* (Figure 5-14A). Transcription profiling during lysogeny for all Cluster A phages tested reveals that expression initiates from P_{rep} and extends through *rep* and across the adjacent genes (Figure 5-5B). The functions of these genes are not known, but the expression data suggest that they may be involved in immunity, and CRS repressor-mediated immunity may be improved if a larger segment of the repressor locus is used.

I examined the impact of these repressor-adjacent genes on immunity using the DaVinci immunity system. A DaVinci lysogen confers complete immunity against Gladiator, DaVinci, phiTM46, and EagleEye, but a CRS carrying an extrachromosomal plasmid with the 1.5 kb DaVinci *rep* locus (pTM48) confers slightly weakened immunity to these phages (Figure 5-30A, B). I constructed a CRS that carries an extrachromosomal plasmid (pTM51) with a 3 kb segment derived from DaVinci that extends from *rep* to gene 73 (Figure 5-30A). The three additional genes are not highly expressed during lysogeny, and they do not have a known function, although gene 74 exhibits sequence similarity to a Cas4-family exonuclease and is conserved in all Cluster A phages. In contrast to pTM48, pTM51 confers complete immunity to phages, mirroring the lysogen profile (Figure 5-30B). Immunity could be strengthened in several ways. One of the three downstream genes may be sufficiently expressed, and may impact repressor

stability, specificity, or abundance. Alternatively, transcription or translation across the repressor gene may be increased or stabilized by the presence of downstream regulatory signals.

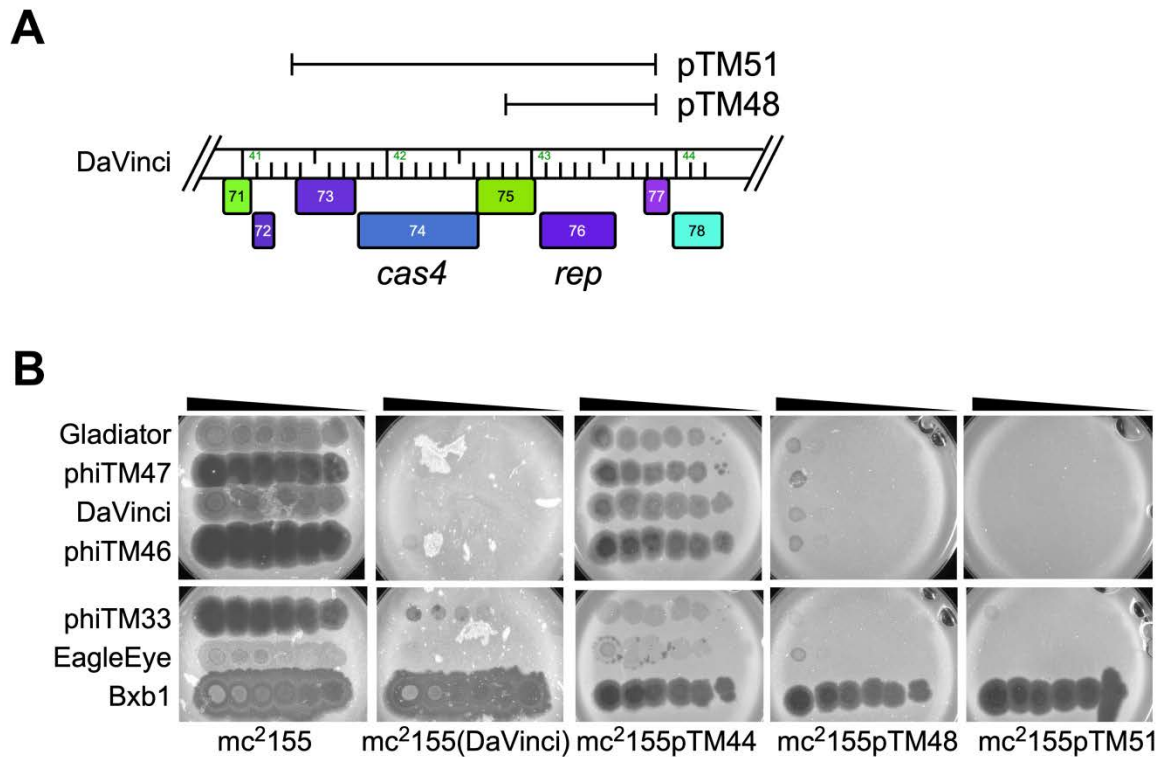


Figure 5-30. Extended DaVinci *rep-73* construct enhances immunity.

(A) Enlarged view of DaVinci immunity repressor locus with the *rep* and *rep-73* regions cloned into pTM48 and pTM51 indicated. (B) Representative immunity assays comparing infection phenotypes of several phages against mc²155, a DaVinci lysogen, and CRSs (pTM44: empty vector; pTM48: DaVinci *rep*; pTM51: DaVinci *rep-73*). Bxb1 is used as a heterotypic control.

5.3.10 The un-regulated extended repressor locus is toxic

Initial characterization of L5 mutants that escaped immunity of an L5 CRS revealed that it was common for the immunity repressor to acquire nonsense or frameshift mutations that truncate the gene product downstream of the helix-turn-helix domain (Donnelly-Wu et al., 1993). From these observations, it was hypothesized that the truncated N-terminal repressor product conferred a dominant negative phenotype by recognizing binding sites without inhibiting transcription. In this model, the truncated repressor outcompetes the full-length repressor expressed from the prophage or plasmid and prevents repressor-mediated defense. Many Cluster A defense escape mutants have acquired similar mutations, suggesting this is a common characteristic of this immunity system (Figure 5-16D).

However, this model may not be correct for several reasons. First, the frameshift that occurs in the phiTM46 repressor that enables escape from a Gladiator CRS occurs within the helix-turn-helix domain, so the C-terminally truncated gene product is not expected to bind DNA (Figure 5-16D). Second, since phiTM38 is a derivative of D29, it has already completely lost the helix-turn-helix DNA-binding domain through a deletion event, yet it gains virulence against a StarStuff lysogen only after acquiring a point mutation in the 3' end of the repressor gene remnant (Figures 5-16A, 5-21A). Interestingly, the mutation introduces a nonsense mutation in the repressor gene's translational frame, truncating the C-terminal portion of the gene if it is expressed (Figure 5-16D). Third, investigation of the CI repressor from the lambdoid phage 434 has revealed that the loss of the C-terminal domain substantially impacts binding specificity of the truncated gene product (Carlson and Koudelka, 1994). Since the Cluster A repressor is structured similarly to CI, loss of the C-terminal domain may have a similar impact.

I directly tested the impact of a truncated Cluster A repressor gene product on immunity using phiTM46, a DaVinci DEM. I created an extrachromosomal plasmid construct carrying the phiTM46 repressor locus (pTM53), identical to the DaVinci repressor construct (pTM48) except for the 1 bp insertion. A Gladiator lysogen is completely immune to DaVinci, but not to phiTM46 or phiTM33, a lytic mutant derivative of Che12 in which the 5' end of the repressor gene has been lost due to a deletion event (Figures 5-16D, 5-31A). A DaVinci lysogen is immune to DaVinci and phiTM46, and partially immune to phiTM33 (Figure 5-31A). A CRS carrying an integrated empty vector (pMH94) and pTM48 is immune to DaVinci, phiTM46, and phiTM33 similar to a DaVinci lysogen, but a CRS carrying pMH94 and pTM53 is not immune to DaVinci or phiTM46, similar to a CRS control strain carrying pMH94 and an extrachromosomal empty vector (pTM44) (Figure 5-31B). However, immunity of a CRS carrying an integrated Gladiator repressor (pTM34) and pTM53 phenotypically mirrors a CRS carrying pTM34 and pTM44, indicating that the presence of the truncated phiTM46 repressor does not interfere with Rep_{Gladiator}'s ability to block infection (Figure 5-31B). Additionally, no phenotypic differences are observed when pTM53 is present in a Gladiator or DaVinci lysogen (Figure 5-31C).

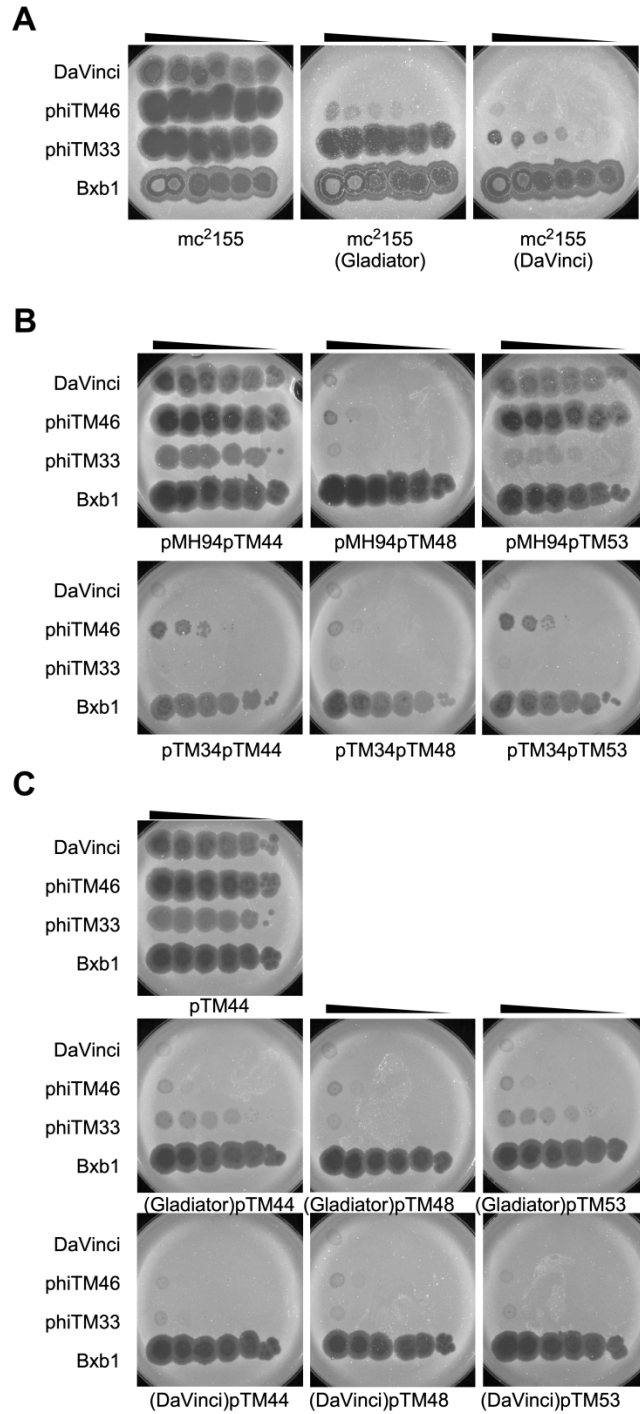


Figure 5-31. phiTM46 *rep* does not confer a dominant negative phenotype.

(A) Immunity assays involving DaVinci and Gladiator lysogens. (B) Immunity assays as in panel A using CRSs with integrated constructs (pMH94: empty vector; pTM34: Gladiator *rep*) and extrachromosomal constructs (pTM44: empty vector; pTM48: DaVinci *rep*; pTM53: phiTM46 *rep*). (C) Immunity assays as in panel A using mc²155 or lysogens (Gladiator, DaVinci) carrying cloned extrachromosomal constructs as in panel B.

Since truncation of the immunity repressor does not appear to directly interfere with immunity, it may impact expression of other genetic factors, such as the genes immediately downstream of *rep* that enhance immunity (Figure 5-30). I constructed a plasmid carrying the extended *rep-73* construct from phiTM46 analogous to the DaVinci segment in pTM51 (Figures 5-30A, 5-32). Although this plasmid, pTM54, can transform *E. coli* without difficulty, it is not able to be efficiently electroporated into either wild type *M. smegmatis* mc²155, or *M. smegmatis* mc²155 carrying the integrated empty vector pMH94 (Figure 5-32). Additionally, pTM54 is unable to be transformed into a Gladiator CRS, a Gladiator lysogen, or a DaVinci lysogen (Figure 5-32). For some strains tested, colonies do eventually appear on the transformation plate, but compared to the empty vector positive control they are very tiny, suggesting that the colonies have acquired genetic mutations conferring antibiotic resistance instead of reflecting true transformants. In contrast, these strains are able to be transformed with pTM51, expressing the full length Rep_{DaVinci} (Figure 5-32). These transformation results suggest that expression of genes in this construct are toxic to *M. smegmatis*, and that the Gladiator and DaVinci repressors expressed in the CRS or lysogens are unable to efficiently regulate the expression of these genes *in trans*.

Furthermore, no strains can be transformed with pTM58, a construct derived from pTM54 in which *rep* has been completely removed, indicating that the truncated repressor product itself is not interfering with transformation efficiency (Figure 5-32). Most phages in the L5 clade contain stop operators in the *rep* promoter, suggesting it is auto-regulated. The inability to transform pTM54 and pTM58 into *M. smegmatis* can be explained by strong toxicity of genes 73, 74, or 75 that are overexpressed from the *rep* promoter in the absence of Rep. The endogenous, auto-regulated immunity repressor expressed in the Gladiator CRS, Gladiator

lysogen, or DaVinci lysogen, may not be present at sufficient levels to regulate the exogenous repressor promoter in these constructs. The expression of these genes during superinfection may promote lytic growth, enabling the phage to overcome the lysogen's or CRS's immunity system.

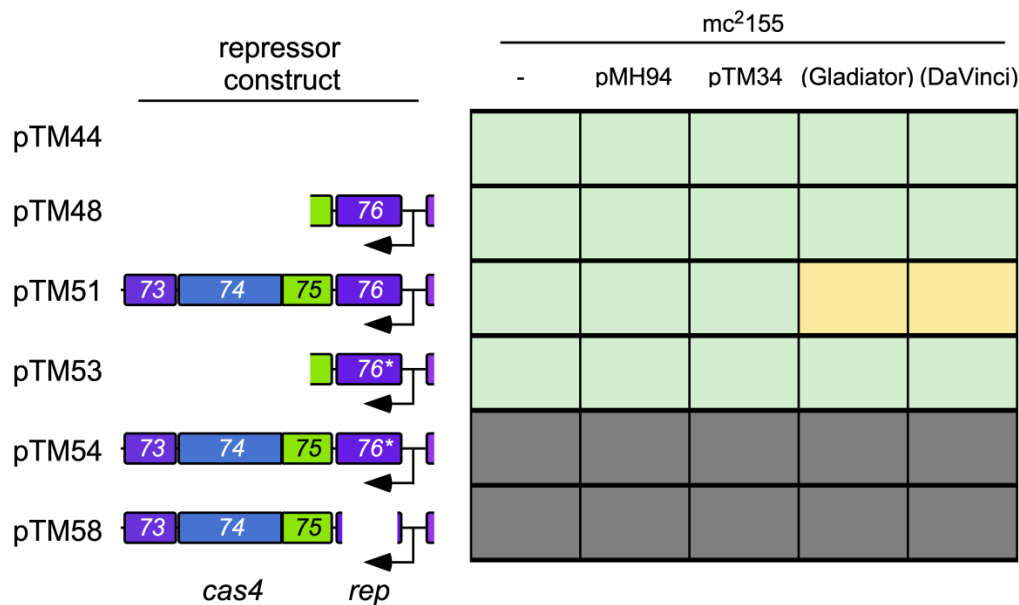


Figure 5-32. Extended phiTM46 repressor construct cannot be transformed.

(left) Diagram of plasmid constructs indicating cloned segments of DaVinci repressor locus, with gene names labeled and direction of transcription indicated with an arrow. Star indicates the 1 bp insertion in present in phiTM46 *rep*. **(right)** Heatmap of transformation results for each plasmid into several strains, including mc²155, lysogens (Gladiator, DaVinci), or CRSs (pMH94: empty vector; pTM34: Gladiator *rep*). Transformations were scored after 4-5 days. Green = transformants were recovered; yellow = transformants were recovered but grew slow; grey = no transformants recovered.

5.3.11 A Bxb1 DEM escapes a Bxb1 CRS

Bxb1 harbors an immunity system very similar to L5, even though it is grouped into Subcluster A1 (Jain and Hatfull, 2000). Similar to L5 clade phages, a CRS carrying the Bxb1 repressor confers immunity to Bxb1 (Figure 5-33A). Immunity is not complete though, and a Bxb1 defense escape mutant, phiTM45, was isolated and purified. Dan Russell and Rebecca Garlena sequenced and assembled this mutant genome. Similar to L5 clade phage DEMs, this Bxb1 derivative has acquired a nonsense mutation in *rep* (Figure 5-33B). This DEM is able to escape Bxb1 CRS immunity at an efficiency of plating of 1, although growth appears to be inhibited in two ways. First, plaques are noticeably smaller on mc²155pTM32 compared to mc²155pMH94. Second, plaques do not appear until 48 h after plating, a phenotype not observed by infection of the heterotypic phage Gladiator (Figure 5-33A). Although the molecular basis of the delayed phenotype is not evident, it suggests that although the *rep* nonsense mutation enables phiTM45 to overcome transcriptional regulation by Rep_{Bxb1}, growth is substantially inhibited and delayed. Delayed superinfection phenotypes have been observed among L5 clade phages but were not specifically observed among DEMs isolated against L5 clade CRSs. The infection pattern of phiTM45 thus clearly indicates that mechanisms of Subcluster A1 immunity system evolution may occur similarly as in L5 clade phages, and it also indicates that delayed superinfection may be directly related to molecular interactions associated with elements of the two immunity systems.

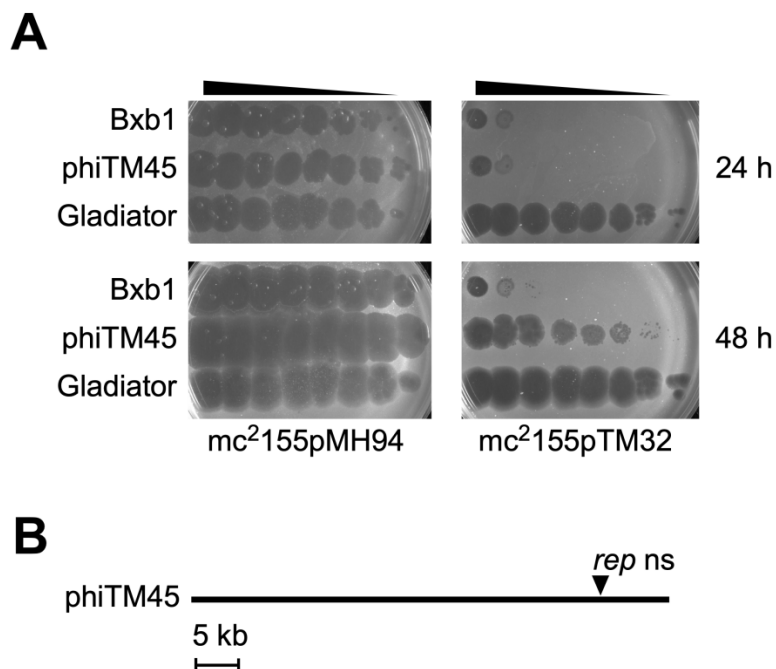


Figure 5-33. phiTM45 DEM exhibits delayed superinfection of a Bxb1 CRS.

(A) Immunity assay of Bxb1 and phiTM45 on CRSs (pMH94: empty vector; pTM32: Bxb1), photographed after 24 h and 48 h. Gladiator serves as a heterotypic control. **(B)** Genome map of phiTM45, a Bxb1 DEM that has escaped a Bxb1 CRS, labeled as in Figure 5-16A.

5.4 DISCUSSION

Classical models of superinfection, immunity, virulence, and the evolution of new immune specificities were primarily developed with limited collections of enterobacteria phages related to λ , P22, and P1 (Campbell, 1994; Yarmolinsky, 2004). However, the genetic diversity among Cluster A phages illustrates how an immunity system can gradually evolve into homologous, mesotypic circuits with regulatory elements that exhibit a spectrum of interactions. Interactions between mesotypic systems generate superinfection immunity patterns that are not necessarily binary or reciprocal. In this more complex genetic environment, virulence and

specificity can be shaped by both homotypic and mesotypic phages. For instance, the mutations acquired in phiTM38 and phiTM39 conferring escape from mesotypic immunity systems also enable homotypic virulence. Additionally, there are different degrees of virulence within a mesoimmunity group. Both phiTM38 and phiTM4 exhibit strong homotypic virulence with little to no impact on escape from mesotypic systems. phiTM39 exhibits weak homotypic virulence but slightly enhanced mesotypic virulence. phiTM42 exhibits both strong homotypic and strong mesotypic virulence.

Homotypic virulence can occur by disrupting interactions between the repressor and cognate binding sites, as observed for λ and P22 (Campbell, 1994), or by disrupting interactions between other factors involved in the regulatory circuit, as seen in P1 and P7 (Heinrich et al., 1995). The Cluster A Rep is the only identified transcriptional regulator involved in initiating and maintaining lysogen (Donnelly-Wu et al., 1993). However, we do not fully understand how the Cluster A immunity system functions, and other factors may be involved, similar to the P1/P7 system. The diverse types of mutations in phiTM39, phiTM40, phiTM41, and phiTM42 conferring escape from a Trixie lysogen may be targeting different aspects of the immunity system. phiTM39 and phiTM40 escape with a modified Rep or modified Holin (or both), while phiTM41 escapes with a modified gp89. The dramatic recombination in phiTM42 enabling escape from all lysogens may combine discordant regulatory elements in Trixie and RedRock that no individual prophage is able to properly regulate at the same time. Meanwhile, the mutation in phiTM4 appears to disrupt a very specific interaction present in L5, L5 derivatives, and StarStuff, such that it does not disrupt interactions in other prophages.

Strains harboring a cloned repressor exhibit stronger defense than the corresponding lysogen for some, but not all phages. This could be caused by several factors. First, expression of

the cloned repressor could be higher than in the corresponding lysogen, and increased expression could compensate for weaker binding affinity to successfully block infection. Alternatively, there could be other factors expressed in the prophage, such as the adjacent downstream gene of unknown function, that negatively impact Rep binding. With the absence of this factor in the CRS, Rep has stronger affinity for stoperators and can block infection.

A more detailed examination of how the Cluster A immunity system functions to maintain lysogeny and prevent superinfection could enhance our understanding of how it evolves. We do not understand which stoperators and operators are used during lysogeny and superinfection. The contribution of individual sites can be measured by performing ChIP-seq with lysogens harboring phiTM1 or phiTM6 prophages (expressing the tagged repressor). Additionally, the precise binding affinities of Cluster A immunity repressors can be determined and compared to each other *in vitro* using high throughput strategies, such as the Bind-n-Seq technique, in which purified repressors are incubated with a library of oligonucleotides to determine binding specificity and affinity (Zykovich et al., 2009). The contribution of individual binding sites, such as the empirically determined operators in L5 (Brown et al., 1997) and Bxb1 (Mediavilla et al., 2000), can be measured *in vivo* by an operator conversion experiment. Using BRED (Marinelli et al., 2008), the operator in L5 can be mutated to the Bxb1 operator, and this mutant can be tested for lysogenization efficiency, the ability of the lysogen to defend against wild type L5, and the ability for the mutant to superinfect a wild type L5 lysogen or a wild type Bxb1 lysogen. Lastly, some Subcluster A3 phages appear to carry an immunity repressor that is more closely related to Subcluster A4 phages than to the other Subcluster A3 phages (data not shown); analysis of this subset of phages may reveal factors associated with how the repressor co-evolves with stoperators.

Mesoimmunity groups are likely to be common. Several groups of actinobacteriophages infecting *Gordonia*, *Rhodococcus*, and *Streptomyces* hosts harbor immunity systems similar to the Cluster A *Mycobacterium* phages (Pope et al., 2017; Smith et al., 2013). Additionally, isolated examples of asymmetric and incomplete infection have been observed among phages related to λ (Kameyama et al., 1999) and P2 (Karlsson et al., 2006), suggesting that the evolution of these heterotypic immunity systems generate similar immunity patterns. A detailed investigation of closely related temperate phages from other clusters utilizing different types of immunity systems could provide deeper insight into patterns of immunity system evolution. However, the Cluster A L5 clade phages evolve within the low gene content flux evolutionary mode (Chapter 2). Investigation of immunity system evolution among temperate phages that evolve within the high gene content flux mode may be more complicated: since these phages exhibit higher levels of horizontal gene transfer, they may acquire diverse phage defense genes, obscuring the contribution of immunity systems to prevent superinfection.

From these analyses, several hypotheses can be generated regarding how the Cluster A immunity system functions. The sequence diversity observed between homologous stoperators near P_{left} in Subclusters A6 and A9, as well as the variability in expression among Subcluster A9 phages Alma and Pioneer, suggest that expression and Rep-mediated regulation at P_{left} may be dynamic. There may be multiple promoters, and Rep may be able to control expression from any of them using different stoperators. The non-coding region immediately downstream of P_{left} is highly expressed in many phages, and this region is void of stoperators, suggesting it plays an important role in lytic growth, perhaps by initiating transcription-dependent replication (Weigel and Seitz, 2006). The first gene downstream of P_{left} may also be important in regulating lysis versus lysogeny, since the L5 derivative, phiTM41, is able to escape Trixie immunity after

acquiring a mutation in gene 89. The two genes immediately downstream of *rep* may also play a role in regulating lysis versus lysogeny. The *cas4-family* gene is one of the few genes completely conserved in all Cluster A phages, and it may be important in replication (Weigel and Seitz, 2006). Interference of *rep* expression may lead to increased expression of the *cas4-family* gene, replication, and lytic growth. The gene immediately downstream of *rep* or immediately downstream of P_{left} may encode an anti-repressor that directly interferes with Rep, similar to Ant in P22 or Coi in P1. In this case, the anti-repressor specifically interacts with Rep, and mutations in the anti-repressor can increase activity to enable escape from defense. Expression of the two genes downstream of *rep* may be regulated at the translational level. Many escape mutants have acquired mutations that truncate *rep*, and the most striking mutant is the D29-derivative, phiTM38, which already lacks the 5' end of the *rep* gene. In many Cluster A phages there appears to be a gap between the *rep* coding sequence and the adjacent downstream gene, and there may be transcriptional or translation signals in this sequence that are impacted by translation of *rep*. These different aspects of the Cluster A regulatory circuitry need to be investigated in order to fully understand how it controls lysogeny and immunity.

6.0 CONCLUSIONS

6.1 SUMMARY OF RESULTS

Since temperate phages can remain in the host as a latent prophage, they encounter unique evolutionary challenges and benefits compared to obligately lytic phages. They must carry a genetic system ensuring their genome is stably inherited as an integrated or extrachromosomal prophage as well as a genetic system ensuring lytic genes are repressed. Although latency may provide safe harbor from non-ideal extracellular environments and broader access to the phage gene pool, it also exposes the phage to deactivation through host-derived mechanisms or through superinfection. My investigation of temperate actinobacteriophages has expanded our understanding of the diverse ways temperate phages evolve within this context.

6.1.1 Whole genome evolutionary patterns

In Chapter 2, I have refined our perspective on how phages evolve at the level of their whole genome. Phages exhibit two distinct evolutionary modes characterized by the degree of gene content flux they exhibit, which differ by an order of magnitude. Groups of genetically related phages generally evolve within one mode or the other, and the mode varies by the type of phage, the type of host, and the phage lifestyle. Although obligately lytic phages evolve within one of the modes, temperate phages evolve within both modes. As a result, temperate phages can be divided into two classes depending on their evolutionary mode: those in Class 1 exhibit high

gene content flux and those in Class 2 exhibit low gene content flux. It is interesting that many factors that are commonly associated with temperate phages and that have distinguished them from obligately lytic phages [such as the lambdoid phages that established the paradigms of phage mosaicism and evolution (Highton et al., 1990; Juhala et al., 2000), and the temperate phages that impact bacterial pathogenicity (Brussow et al., 2004)] tend to be derived from temperate phages grouped in Class 1. Nearly half of all temperate phages are grouped in Class 2 and evolve much more similarly to obligately lytic phages. It is also interesting that the distribution of phages within the two evolutionary modes differs by host phylum. In hosts of Actinobacteria and Proteobacteria, both classes of temperate phages are identified. In contrast, in hosts of Cyanobacteria only Class 2 temperate phages are identified, and in hosts of Firmicutes temperate phages are predominantly Class 1.

We fundamentally do not know what causes these evolutionary modes. The two modes could be driven by different absolute levels of HGT or by different sized gene pools that phages have access to during HGT (Hendrix et al., 1999; Jacobs-Sera et al., 2012). Additionally, the different proportions of phages in each mode observed between host phyla could represent biases in phage isolation techniques; within each host phylum, certain subsets of phages are more easily isolated than others, misrepresenting the complete diversity. Furthermore, since the evolutionary modes were examined in closer detail using manually curated groups of actinobacteriophages, the analyses should be extended to groups of genetically-related phages infecting other host phyla to determine if they also exhibit a correlation between evolutionary mode (determined from a genomic similarity plot) and HGT (determined from a phylogenetic tree). Since actinobacteriophage Cluster A contains phages representing both evolutionary modes, it provides model phages to empirically examine the biological basis of the two modes, since they control

for many genetic variables such as genome architecture, host, immunity system, prophage inheritance system, etc. Last, we do not know whether temperate phages of different evolutionary modes impact their environment differently. Analyses of temperate phages isolated from specific environments, such as the gut microbiome (Kim and Bae, 2018) or the marine (Knowles et al., 2016) environment, could provide further insight.

6.1.2 *Bifidobacterium* prophages

In Chapter 3, I have expanded our understanding of temperate phage diversity by characterizing and inducing prophages of *Bifidobacterium* hosts. There are approximately 130 genera within the phylum Actinobacteria (Barka et al., 2016; Ventura et al., 2007), but phages have been isolated from hosts in only 14 genera, so this study provides insight into the genetic diversity in phages infecting unexplored genera. Prophages integrated at the tRNA^{Met} locus likely utilize an integration-dependent immunity system homologous to systems in *Mycobacterium* phages such as BPs (Broussard et al., 2013). Prophages integrated at the *dnaJ₂* locus rely on a tyrosine integrase, which is not common when the host integration site is positioned within a coding sequence (Hatfull, 2012). Furthermore, the use of *dnaJ₂* as an integration site is interesting since this gene is specific to hosts in Actinobacteria (Ventura et al., 2005a).

This analysis also suggests that many bifidophages utilize phase variation systems to modulate their host specificity, which is interesting for several reasons. The most highly characterized phage-related systems, including Min, Cin, and Gin, are derived from enterobacteria phages (Sandmeier, 1994). The bifidophage systems are the first such systems reported in phages infecting Actinobacteria. Furthermore, Min, Cin, and Gin all rely on serine-family recombinases in *cis* (Johnson, 2015). In contrast, bifidophages integrated at tRNA^{Met}

may utilize a simple inversion system that relies on a host recombinase in *trans*, and bifidophages integrated at *dnaJ₂* may utilize a more complex shufflon, Rin, that relies on a tyrosine-family recombinase in *cis*.

Although I was able to successfully induce at least two of the prophages to form complete phage particles, I was unable to confirm they are infectious. There may be many *bona fide* technical or biological reasons why infection was not observed, but I cannot rule out that the prophages are simply defective and cryptic. Further optimization and testing of media, experimental conditions, and indicator strains are needed to determine which of these hypotheses are true. The bifidobacterial lysogens I have characterized, along with the protocols established for prophage induction, provide a model system to further explore why no infectious bifidophages have been yet identified. Similarly, although I have bioinformatically characterized two potential phase variation systems, I do not have *in vivo* or *in vitro* evidence that they impact host range specificity. Further work needs to be done to test whether the RBP alleles expressed from these loci exhibit distinct biochemical specificities.

6.1.3 Partitioning systems

In Chapter 4, I have helped to generate the first characterization of actinobacteriophage partitioning systems. Detailed analyses of enterobacteria phages P1, P7, and N15 have shaped the paradigms for how some prophages avoid integrating into the host chromosome during lysogeny (Abeles et al., 1985; Hayes and Austin, 1993; Lobocka et al., 2004). However, few other prophage-related partitioning systems have been characterized. I show that extrachromosomal Cluster A *Mycobacterium* prophages utilize *parABS* systems that are structured similarly to previously described plasmid-related Type Ib systems (Ebersbach and

Gerdes, 2005; Schumacher, 2012). Similar to the effects that partitioning systems have on plasmids, the *parABS* systems confer partition-mediated incompatibility between co-resident extrachromosomal prophages, and these phages may rely on two evolutionary strategies to avoid this problem. First, the *parABS* cassettes may possibly be exchanged with *parABS* cassettes that carry different incompatibility determinants or with integration cassettes. Second, unlike *parA*, *parB* exhibits relaxed selection, suggesting that it may evolve new specificities in conjunction with *parS* sites to avoid genetic interactions with other systems.

Several experiments are needed to further examine the actinobacteriophage partitioning systems. To identify whether they represent multiple compatibility groups, binding assays using additional purified ParB homologs can be performed, complemented by compatibility experiments using plasmids carrying different cloned partitioning systems. To further confirm that the *Mycobacterium* partitioning systems function similar to other Type Ib systems, ParA and ParB could be fluorescently labeled to assess where they localize during host growth. We do not understand the selective pressures for different prophage inheritance strategies. Cluster A phages suggest that exchange of modular integration and partitioning systems is possible. To determine the extent to which *parABS* and integration cassettes perform completely alternative functions, recombinant phages can be engineered using highly characterized integrating phages (such as L5) to carry a partitioning system instead of an integration system, and their stability during lysogeny and impact on the host can be directly compared to the wild type integrating phage.

6.1.4 Immunity systems

In Chapter 5, I have refined our understanding of the relationship between immunity systems and superinfection immunity. Paradigms of how immunity systems function and evolve,

how phages with genetically related immunity systems form distinct immunity groups, and how phages may acquire virulent escape from immunity, were established using phages related to λ (Ptashne, 1992), P22 (Susskind and Botstein, 1978), and P1 (Heinrich et al., 1995). However, these analyses were limited since they did not have available a large number of genetically related phages spanning a spectrum of genetic diversity with completely sequenced genomes. Now, using the large collection of well-defined and closely-related Cluster A phages, I show that evolving immunity systems result in asymmetric and incomplete immunity phenotypes.

Phages with a genetic spectrum of immunity systems do not form clearly-defined immunity groups, but instead form a mesoimmunity group in which any two phages may exhibit homoimmunity, heteroimmunity, or mesoimmunity. Mesoimmunity groups also impact the ability for phages to gain virulence. Mutations may confer virulence against some, but not all, immunity systems within the mesoimmunity group. As a result, Cluster A phages may not be able to gain virulence against homotypic prophages, but they are able to gain virulence against mesotypic prophages, which may drive diversification of immune specificities. This also means that divergence of immune specificities may take an indirect path from homoimmunity towards heteroimmunity, such that immunity phenotypes between phages are dynamic and frequently changing.

Although mesoimmunity is likely to be common among phages with different types of immunity systems, more work needs to be done to evaluate this. The analysis in Chapter 5 focused on 100 phages related to L5. There are many more phages in Cluster A, and their diversity can be investigated to determine if they exhibit mesoimmunity as well. Some groups of λ -related and P2-related phages may exhibit mesoimmunity, but the reports were more anecdotal than systematic (Kameyama et al., 1999; Karlsson et al., 2006). Temperate *Mycobacterium*

phages in Clusters F and K are abundant, genetically diverse, and unrelated to Cluster A phages, so they would be good candidates to extend this analysis (Figure 2-6C). Additionally, since the Cluster A immunity system is not fully characterized, further work needs to be done to understand which genetic elements are directly contributing to the observed immunity phenotypes. The DEM series highlights additional genetic loci that may be involved, and these can be examined. Many *in vivo* and *in vitro* experiments can also be conducted to identify which stopoperator sites within the prophage are used by the cognate Rep to maintain lysogeny and which sites are targeted by Rep from other prophages during superinfection. Although the majority of Cluster A phages infect *Mycobacterium*, Subcluster A15 phages infect *Gordonia* (Pope et al., 2017). Comparative analyses between Cluster A phages of each host can highlight how this immunity system evolves as phages change hosts.

6.2 FUTURE DIRECTIONS

6.2.1 Dynamics of lysogeny

We do not understand the ecological pressures imposed on temperate phages. It is common and straightforward to isolate mutant derivatives of temperate phages, in the absence of selection, that are no longer able to lysogenize the host (Campbell, 1994; Donnelly-Wu et al., 1993). The frequency by which these obligately lytic mutants are generated suggest there is strong selective pressure to maintain the ability to lysogenize the host in the natural environment, but we do not fully understand these pressures. Specifically within the host phylum *Mycobacterium*, many phages are temperate, but reliable strategies to artificially induce lytic

growth during lysogeny have not been identified. Although induction strategies, such as the use of mitomycin C, have been effective for lambdoid prophages (Rokney et al., 2008), *Lactococcus* prophages (Oliveira et al., 2017), and even *Bifidobacterium* prophages (Chapter 3), these strategies have not worked well for *Mycobacterium* prophages (Hatfull, 2012). This contrast suggests that *Mycobacterium* prophages do not rely on the host SOS response for induction, but we do not understand how they are monitoring the intracellular environment. Unlike other hosts such as enterobacteria or *Bifidobacterium*, the majority of sequenced *Mycobacterium* genomes do not contain any prophages (Fan et al., 2014), possibly suggesting a connection to the observed challenges with induction.

6.2.2 Tracking host history

Hosts used to isolate or propagate a phage in a laboratory setting are not necessarily the optimal hosts phages utilize in the natural environment. The frequency by which phages switch hosts, and the genetic distance between the hosts, are not well understood. *Mycobacterium* phage Patience (Pope et al., 2014) and enterobacteria phages BP-4795 and cdtI (Chithambaram et al., 2014b) provide examples of host switching, but these have been identified due to large contrasts in sequence adaptations. Switching between closely-related hosts, which is likely to be more common, is not as easily measured this way. Currently, we are not able to identify the optimal host or historical host for any given temperate phage, but improvements in sequencing technology may enable us to robustly pair phages with hosts in their environment and understand this more clearly. First, metagenomic analyses are able to identify and characterize phage diversity within environmental samples in an unbiased way (Breitbart et al., 2002; Edwards and Rohwer, 2005), which may help to identify novel temperate phages that are not able to be

isolated using common laboratory techniques. Second, using single-cell sequencing, prophage genomes can be precisely matched to the host in which they are residing within the environment (Roux et al., 2014). Third, longer sequencing read lengths enable sequencing of the entire phage genome derived from an individual virion particle instead of from large populations in a prepared lysate (Allen et al., 2011; Klumpp et al., 2012), improving our ability to probe phage diversity within environments. When we have a large dataset connecting diverse temperate phages to their hosts, we can conduct detailed analyses of how often they switch hosts, as well as potentially be able to predict optimal hosts for lysogeny.

6.2.3 Tracking horizontal gene transfer

It has long been recognized that phages exhibit substantial levels of horizontal gene transfer, and a phage may participate in this transaction with its host, with other co-infecting phages, or with other genetic elements such as plasmids. However, we do not fully understand the dynamics of these processes. The frequency or direction of HGT for any particular gene or groups of genes may be dependent on the specific type of donor and recipient genomes. For instance, phages identified through metagenomic analyses of the gut microbiome have suggested that the abundance of phages from diverse hosts enhances the transfer of gene flow across larger genetic distances (Lugli et al., 2016b). As large collections of phages are isolated and sequenced from large collections of genetically diverse and genetically related hosts, we may be able to develop a more detailed map of HGT. The types of genes that benefit from this can be identified, the direction of flow and frequency of HGT across phyla, genera, and species can be quantified, and the roles that temperate phages play can be assessed.

6.2.4 High throughput phage genomics

The large collection of actinobacteriophages that has been generated through the SEA-PHAGES program provides an unparalleled resource to investigate temperate phage diversity and evolution. However, leveraging the genetic diversity in this collection can be challenging. For instance, the large number of superinfection immunity assays described in Chapter 5 were performed manually and were very time intensive, limiting the interrogation to only 32 strains and 45 superinfecting phages. In order to fully utilize the genetic diversity among hundreds of Cluster A phages specifically, as well as among the entire phage collection in general, development of high throughput techniques to screen through phages with automated or semi-automated strategies would be advantageous, similar to what has been done in *Acinetobacter* phages (Schooley et al., 2017).

6.2.5 Biotechnological applications

The insights derived from my research can be used to develop several new genetic tools for synthetic biology and to study bacterial hosts. The characterization of actinobacterial partitioning systems can aid in the development of extrachromosomal vectors for *Mycobacterium*, *Gordonia*, and *Streptomyces* hosts. The partitioning systems in *Mycobacterium* phages may represent three distinct compatibility groups, with Echild and 40AC carrying ParB genes most divergent from the rest. The characterization of the tyrosine-family *dnaJ*₂-targeted integration system can be used to develop integration vectors specific to bacteria in the phylum Actinobacteria. The characterization of the Cluster A immunity repressors may be useful for creating artificial biological circuits. Some of these repressors exhibit orthogonal binding

specificities that can be used in parallel circuits while other repressors exhibit a spectrum of binding specificities that can be used to develop asymmetric or graded circuits. Similarly, the bifidophages simple inversion system in tRNA^{Met}-integrated prophages could enable unique, binary, *trans*-regulated circuitry while the more complex shufflon in *dnaJ*₂-integrated prophages could enable unique *cis*-regulated non-binary circuitry that could generate combinatorial or stochastic variation.

6.2.6 Evolution of viral latency

The evolutionary benefits that temperate phages gain through latency suggest that viruses of other host domains would also derive similar benefits. Fewer than 100 viruses infecting hosts in the domain Archaea have been reported (Zhang et al., 2012b), so little is known about their diversity. However, temperate archaeal viruses are likely to be abundant and diverse, since several have already been shown to form lysogens, including ΦCh1 (Iro et al., 2007) and SNJ1, which can be induced with mitomycin C (Zhang et al., 2012b). Interestingly, SNJ1 remains as an extrachromosomal provirus during lysogeny (Zhang et al., 2012b), suggesting it may utilize partitioning systems similar to extrachromosomal phages. In contrast, among the thousands of viruses identified that infect hosts in domain Eukarya, only a few, such as herpesvirus, exhibit latency similar to temperate phages (Kuhn et al., 2019; Speck and Ganem, 2010). Herpesviruses remain as circularized plasmids in the host nucleus and monitor the intracellular environment. Similar to lysogeny, the latency is controllable and reversible, and herpesviruses initiate a cascading transcriptional program after induction. Overall, the evolutionary benefits of latency may become diminished for viruses infecting Eukarya compared to viruses infecting Archaea and Bacteria, and future studies may be able to identify this evolutionary progression.

6.3 CODA

Historically, the examination of temperate phage diversity and evolution has yielded many insights into basic biological processes and has generated many biotechnological benefits. This dissertation has advanced our understanding of temperate phages a little bit further. Perhaps phage research in the upcoming 100 years will be as fruitful and impactful as it has been during the first 100 years and bring us even closer to understanding how these diverse biological entities have evolved for billions of years and shaped the world around us.

APPENDIX A PHAMERATOR DATABASE MANAGEMENT

The following chapter describes my contribution to the management of SEA-PHAGES genomics data. Many of the computational and bioinformatic analyses performed in Chapters 2-5 relied on phage genomics data generated by the SEA-PHAGES program. As a result, I have helped to maintain accuracy and integrity of the data by acting as administrator for the database used by Phamerator, which was created by Steve Cresawn. In this role, I maintained and developed the Phamerator database, as well as developed new scripts and improved scripts written by others (including Charles Bowman and Steve Cresawn) to enhance the SEA-PHAGES data management pipeline. In describing my contributions, this chapter also serves as a reference guide for other SEA-PHAGES database administrators and end-users. The work is unpublished, but the scripts are publicly available on GitHub.

A.1 INTRODUCTION

Since the early 2000's, the SEA-PHAGES program has steadily expanded to include thousands of students from over 100 institutions, and it generates hundreds of newly isolated and sequenced phages every year. Not surprisingly, a large volume of genomics data is produced, and a complex pipeline has been steadily constructed to manage this data. The continued success of the program depends on the strategic management and development of this pipeline to

maximize a) the speed and efficiency of data acquisition and processing, b) the accuracy, integrity, consistency, and quality of the data itself, and c) the accessibility of the data for end-users.

SEA-PHAGES genomics data is most valuable if it is able to meet several criteria. First, the genomics data must be accessible to diverse types of end-users, including educators in the classroom, researchers within the actinobacteriophage community, and researchers in the general scientific community. Second, the data needs to be accessible in a simple, straightforward format for any user to quickly retrieve information for particular phages or genes of interest as well as through a command line interface that enables high-throughput computational genomics analysis. Third, the data needs to reflect and facilitate the dynamic, multi-step process of gene annotation as it is generated, reviewed, and refined by various participants of the SEA-PHAGES program. Fourth, the data needs to be integrated with other actinobacteriophage data generated by the general scientific community.

In order to achieve these goals, a data management pipeline has gradually developed. Genomics data are generated and reviewed for quality by many participants and funneled into two databases, PhagesDB and PhameratorDB, that store overlapping, but distinct, types of data. These two databases are directly managed by SEA-PHAGES researchers, are routinely updated, and reflect the entire collection of actinobacteriophage data. This includes SEA-PHAGES genome sequences that are published or unpublished, SEA-PHAGES annotations that are finalized or still in preparation, and genome sequences and annotations derived from researchers that are not associated with the SEA-PHAGES program. Once the iterative, multi-step process of gene annotations is complete, finalized SEA-PHAGES genomes and annotations are deposited in the NCBI GenBank database where they are systematically stored with other (phage and non-

phage) genomics data and are easily accessible to the general scientific community. Although this data is initially identical to data in PhagesDB and PhameratorDB, the SEA-PHAGES team retains only limited control of the data.

PhagesDB and PhameratorDB were created for different purposes. PhagesDB was created by Dan Russell to serve as a comprehensive central source of actinobacteriophage genomics data (including myriad details about each genome, such as where and when it was isolated, who isolated it, etc.) that is easily accessible to end-users through a web interface (Russell and Hatfull, 2017). In contrast, PhameratorDB was designed by Steve Cresawn as an integral part of a larger tool, Phamerator, in order to visualize evolutionary relationships between phages (Cresawn et al., 2011). Phamerator identifies significant nucleotide sequence similarity by whole genome BLAST alignment (Altschul et al., 1990), amino acid sequence similarity by grouping gene products into phamilies (“phams”)(Cresawn et al., 2011), and common protein structural domains using the NCBI conserved domain database (CDD)(Marchler-Bauer et al., 2011). Visualization of these relationships is accomplished by connecting PhameratorDB to a graphical user interface (GUI). Although nucleotide sequence similarity is computed on-the-fly in the GUI, gene product phams and conserved domains are pre-computed and stored in PhameratorDB. Additionally, in contrast to PhagesDB, different instances of PhameratorDB-structured databases can be prepared using different subsets of phage genomes. The primary instance that contains the most up-to-date, complete collection of actinobacteriophage annotation data is the *Actino_Draft* database, but instances containing annotations of phages of particular host genera or of particular annotation sources and stages have been created for specific research projects and can be “frozen” for publication.

The database and GUI that jointly constitute the Phamerator program are separate, independent software entities. The development of computational tools to manage each entity were initiated by Steve Cresawn, but as the number of phage genomes have increased, the processes of PhameratorDB management and GUI software development have also become distinct. Charles Bowman made substantial database management improvements, such as how genomes are imported into the database and how phams are computed. In contrast to the GUI, PhameratorDB can be accessed through a command line interface for high-throughput genomics analyses, and many of my comparative genomics projects have relied on this database. As a result, I too have contributed to the development and management of PhameratorDB. As the database administrator, I have created new scripts, made improvements to previously existing scripts, made minor improvements to the database itself, and ensured the database remains up-to-date and in sync with PhagesDB and GenBank. My efforts to increase automation in the PhameratorDB pipeline have improved data integrity and reliability not only within this database but throughout the entire SEA-PHAGES data pipeline as well.

Here, I provide a summary of the current state of PhameratorDB management, with a special focus on my particular contributions. PhameratorDB management is one of many elements that constitute the broader SEA-PHAGES pipeline, and brief descriptions of other elements are provided when necessary. However, aspects of this pipeline have been described elsewhere (Cresawn et al., 2011; Hanauer et al., 2017; Russell and Hatfull, 2017), and a comprehensive overview of the entire pipeline (and even of PhameratorDB itself), is beyond the scope of this chapter.

A.2 PIPELINE DEVELOPMENT

A.2.1 Overview of SEA-PHAGES data pipeline

PhameratorDB database administration occurs within the context of the entire SEA-PHAGES program, which begins with identification of new phages (Figure A-1). Thousands of undergraduate students isolate and purify new actinobacteriophages in the SEA-PHAGES course (Hanauer et al., 2017). Lysates are sent to the Pittsburgh Bacteriophage Institute Genome Sequencing Center where the DNA is extracted and sequenced, and where the new phage genome is assembled, clustered (and subclustered) relative to other actinobacteriophages, and uploaded to PhagesDB. “Metadata” regarding the genome, including location of isolation, designated cluster and subcluster, and host strain, is added to this database directly from the Genome Sequencing Center or from SEA-PHAGES students.

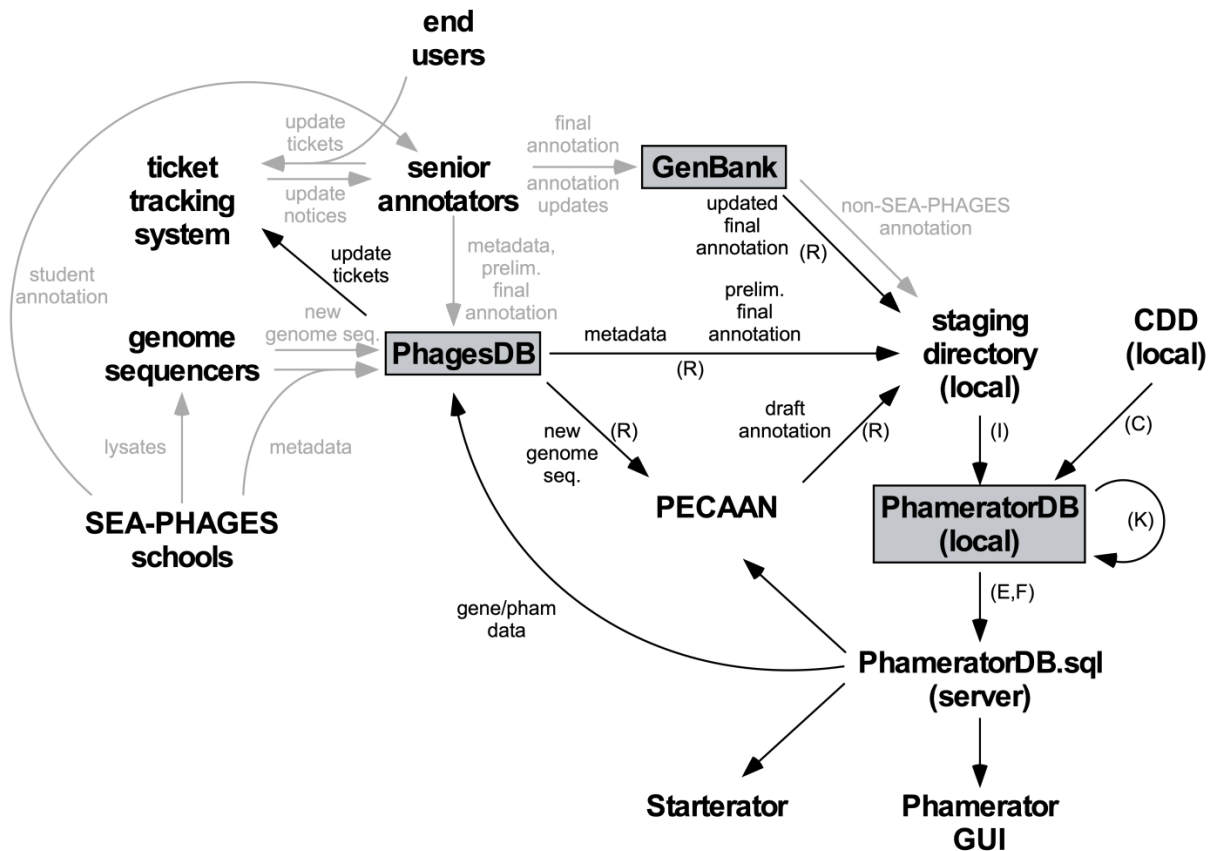


Diagram highlighting the various steps in which phages are isolated, sequenced, annotated, imported into PhameratorDB, and updated in PhameratorDB. Grey lines and text indicate steps that are manually completed. Black lines and text indicate steps that are automatically completed using various scripts. Scripts in the **k_phamerate** repository control several steps of this process, including (R) retrieving data from various sources and systematically staging it in a local directory, (I) importing the retrieved data into PhameratorDB, (C) identifying conserved domain data, (K) computing pham data, (F) freezing databases for publication, and (E) exporting databases to a public server for downstream analysis and applications.

bioinformatics course using a variety of computational tools. A team of trained, senior annotators review *student* annotations and generate *preliminary final* annotations. This annotation set is then evaluated using several automated and semi-automated quality control (QC) checks using PhameratorDB data management scripts that I have developed. If the annotations fail QC, they are revised by annotators and re-submitted as many times as necessary until they pass QC. Once the annotations pass QC, they are accepted into PhameratorDB, and annotators deposit the *final* annotations into GenBank where they are assigned a unique accession identifier. Gene annotations can be continually refined, and corrections or updates can be made to the *final* GenBank annotations as often as needed by authors that are specifically listed in the GenBank record.

Several computational tools rely on PhameratorDB (Figure A-1). PhagesDB relies on PhameratorDB for gene data, including gene coordinates, descriptions, and phams. Starterator, managed by Christopher Shaffer, retrieves data from PhameratorDB and aligns genes within phams to help define gene start coordinates. The Phage Evidence Collection And Annotation Network (PECAAN) website (<https://discover.kbrinsgd.org>) utilizes data in PhameratorDB to provide various tools to help students record and predict gene functions.

A.2.2 Overview of key aspects of PhameratorDB

PhameratorDB is a MySQL database comprised of 13 tables (Cresawn et al., 2011). The *Phage*, *Gene*, *Pham*, and *Version* tables represent the most important tables for the data management pipeline. The *Phage* table contains information that pertains to the entire phage genome, such as the genome sequence, the host strain, the designated cluster, etc. The *Gene* table contains information that pertains to individual genes, including coordinates, orientation, the

gene product, etc. The *Pham* table contains a list of genes from the *Gene* table with their computed pham. The *Version* table keeps track of the database version and is updated every time the database is changed.

The primary PhameratorDB instance, *Actino_Draft*, serves not only as a repository of final, refined gene annotations, but also as a tool to facilitate the dynamic, iterative improvement of annotations. It is therefore valuable to be able to import new *draft* annotation data as soon as possible and rapidly replace it with progressively improving gene annotations. In order to facilitate this process, the pipeline relies on the “GenBank-formatted flat file” as the primary file format for importing annotation data into PhameratorDB (Figure A-2). This is a structured text file that systematically stores diverse types of information about the genome (<https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>). A flat file can be generated for any genome at any annotation stage using GenBank, DNA Master (<http://cobamide2.bio.pitt.edu>), PECAAN, or the python coding package, *Biopython* (Cock et al., 2009). Flat file fields, such as LOCUS, DEFINITION, and REFERENCE-AUTHORS provide information regarding the entire record, while others, such as FEATURES, provide information about particular regions of the sequence in the record, such as tRNA or CDS genes. During the import process, data is parsed from flat files and stored in the *Phage* and *Gene* tables. If a previous annotated version of the genome is already present in the database, all data relating to that genome (including data in the *Phage*, *Gene*, and *Pham* tables) is removed and the genome is completely re-imported using the new flat file. Only in rare circumstances are individual fields in specific tables updated for specific phages with data not directly derived from a flat file.

LOCUS	KF861510	52974 bp	DNA	linear	PHG 05-DEC-2017	FEATURES	Location/Qualifiers
DEFINITION	Mycobacterium phage EagleEye, complete genome.					source	1..52974
ACCESSION	KF861510						/organism="Mycobacterium phage EagleEye"
VERSION	KF861510.1						/mol_type="genomic DNA"
KEYWORDS							/isolation_source="soil"
SOURCE	Mycobacterium phage EagleEye						/db_xref="taxon:1429759"
ORGANISM	Mycobacterium phage EagleEye						/lab_host="Mycobacterium smegmatis mc2 155"
	Viruses; dsDNA viruses, no RNA stage; Caudovirales; Siphoviridae; L5virus; unclassified L5virus.						/country="USA: Houston, TX"
REFERENCE							/lat_lon="29.888211 N 95.462524 W"
AUTHORS	1 (bases 1 to 52974) Awa,H., Bernal,J.T., Coelho,R.E., Culpepper,S.C., Devaraju,V.S., Higgins,R.T., Husein,A.J., Johnston,E.M., Jung,J.A., Kanani-Hendijani,T.A., Knapp,R.E., Lepiocha,N., McCarter,A.J., Merlau,P.R., Monfared,M.S., Olney,H.P., Pineda,M.R., Pizzini,S.E., Roberson,D.J., Rodriguez,J., Simpson,N.A., Stevens,S.C., Stroub-Tahmassi,C.A., Syed,N., Torres,S.E., Townsend,C.W., White,X.E., Willette,C.E., Deming,K.E., Simon,S.E., Benjamin,R.C., Hughes,L.E., Hale,R.H., Lamson-Kim,T., Visi,D.H., Allen,M.S., Bradley,K.W., Clarke,D.O., Lewis,M.F., Barker,L.P., Bailey,C., Asai,D.J., Garber,M.L., Bowman,C.A., Russell,D.A., Pope,W.H., Jacobs-Sera,D., Hendrix,R.W. and Hatfull,G.F.					trRNA	5584..5657 /gene="9" /locus_tag="P81_EAGLEEYE_9" /product="tRNA-Trp" /note="tRNA Trp (cca)"
TITLE	Direct Submission					CDS	complement(45372..45914) /gene="81" /locus_tag="P81_EAGLEEYE_81" /note="homology to L5 Immunity Repressor" /function="DNA binding" /codon_start=1 /transl_table=11 /product="immunity repressor" /protein_id="AHG23861.1" /translation="MSGKTQSGSFRAPLIFSVIEDLRKGYNOSEIADMHGVTQAV SWQKKTYGGRMTPRDIVREAWPFETTLHGKSVFPQRLRDHGFMTGGAGSENKIR RLKAWMRKLDEEDVLEFDPNIPPTPGMAGGGFRVYHRDITDGPDLILRVNEHTAPM TEKSEMIWSFPHDIEILQS"
JOURNAL	Submitted (15-NOV-2013) Pittsburgh Bacteriophage Institute and Department of Biological Sciences, University of Pittsburgh, 365 Crawford Hall, Pittsburgh, PA 15260, USA					ORIGIN	
COMMENT	Phage Isolation, DNA preparation, annotation analysis, and sequencing by Ion Torrent to approximately 90x coverage was performed at University of North Texas, Denton, TX Assembly performed with Newbler and Consed software as of Feb, 2013. Supported by Science Education Alliance, Howard Hughes Medical Institute, Chevy Chase, MD.						1 tgcctgccgaa ccatcggtga cgggttttca agtcgatcag aagaagaggc ctgcactgga 61 aggcctcgga atggcctgt gggccctctt ggctgacgaa caccgcctcc cgcgggtatc 121 tttaccccca aactgcgcag cggccttctg ggcccgttgg ctgtaaaagt gaactacctc 181 acatttacag tgagcacttt gcgatactcc cgtatatata ttatgagggg ctgaaggccc 241 ctctgaagag cgcctttagg gcgctcacta agaactaaag accgcgcttc gagggcgggt 301 catagaacit ggaaccgca acccggggtt ctctggcggc gccagtcgac cgcctaaagg 361 atcggggccc cagtggcgcc ctctaaaggg tgttaactcg tgttggcac cgcctcgaat 421 gtcaactggg acactcaacc ggggaagttc gacgttctga acctcgggat gcggttcgac 481 agctcgtccg agcacgagat ccccgacctg gccgcgacgc acttcgtgcc ggccaacctc 541 gcggcgtgga atatgccgcg acatcgcgaa tacgcgcgca ttccggcgcg cgctctgcac
	##Assembly-Data-START## Assembly Method :: Newbler and Consed v. Feb-2013 Coverage :: 90x Sequencing Technology :: Ion Torrent ##Assembly-Data-END##						

Figure A-2. Parsing of a GenBank-formatted flat file record.

Yellow boxes indicate data that is stored in PhameratorDB. Blue boxes indicate data that is evaluated for quality control, but not stored in PhameratorDB.

The *Actino_Draft* database contains the most up-to-date actinobacteriophage annotation data, and it is routinely updated through a multi-step process (Figure A-1). The most up-to-date version is stored on the database administrator's private, local computer. Data that needs to be added to the database is retrieved from various sources and staged in a structured local directory for import. New data (mostly in the form of flat files) is parsed and imported into the *Phage* and *Gene* tables, conserved domain data is retrieved from a local copy of the NCBI CDD and stored in the *Gene*, *Domain*, and *Gene_Domain* tables, and after genes are grouped into phams the pham data is stored in the *Pham* table. After all updates are implemented, the database version number is updated in the *Version* table. The updated *Actino_Draft* database is exported from MySQL into a single file, **Actino_Draft.sql**, and is uploaded to a public server where it

replaces the old database file. Applications that rely on PhameratorDB data retrieve and parse this new database file.

A.2.3 Modifications to PhameratorDB and data management scripts

Automating updates to PhameratorDB helps to enhance accuracy and data availability for end-users. Therefore, in order to automate several steps in the data management pipeline, I have had to modify the structure of PhameratorDB. First, I have created and modified several fields in the *Phage* table.

PhageID. This field is the primary key of the *Phage* table and is the unique identifier for all phages in the database. Historically, this field had been populated using different strategies, and it had become difficult to correlate phage names in PhagesDB to their *PhageID* in the *Actino_Draft* database. I have revamped this field so that there is a direct correspondence between phage names in PhagesDB and phage names in GenBank records to *PhageIDs* in the *Actino_Draft* database, although with some exceptions.

Name. This field also reflects the phage name, but it is not as constrained as the *PhageID*, and this name is displayed in the Phamerator GUI. For all *draft* genomes, the *Name* contains the *PhageID* with a *_Draft* suffix appended, indicating the annotations have been automatically annotated. For all other genomes, the *Name* corresponds to the *PhageID*.

Accession. This field had not been consistently populated. It is now reliably populated and updated directly from import tickets and is used for auto-updating genomes from GenBank records. It is important to note that the NCBI generates RefSeq records that are derived from GenBank records. After data is submitted to GenBank, authors retain control of the GenBank record but not the RefSeq record. As a result, the PhameratorDB *Accession* field should always

store the GenBank ACCESSION number (and not the RefSeq ACCESSION number) for SEA-PHAGES genomes. For non-SEA-PHAGES genomes, either accession number may be stored. In either case, the *Accession* should not contain the sequence version (represented by the integer to the right of the decimal).

Cluster2 and ***Subcluster2***. Originally, PhameratorDB contained a single field, *Cluster*, that reflects how the phage genome has been grouped relative to other phages in the database, but it combined cluster and subcluster designations. It remained empty (*NULL*) if the phage was a singleton (and was not clustered), was populated with the cluster designation if the phage was clustered (but not subclustered), or was populated with the subcluster designation if the phage was clustered and subclustered. I have now split this data into separate fields: *Cluster2* only contains cluster data, and *Subcluster2* data only contains subcluster data. This primarily facilitates data processing by downstream applications and analyses. The *Cluster* field remains unchanged since certain applications (such as Phamerator GUI) still rely on this field.

DateLastModified. Originally, PhameratorDB did not contain a record of when the genome and annotations were imported. I have created this field with a DATETIME data type so that it can record the date in which a genome and its annotations have been imported. This is valuable to keep track of which annotation version has been imported, and it also facilitates automated updating of the database. It is important to note that the date stored in this field reflects the date the annotation data were imported, and not the date that the annotation data were created.

RetrieveRecord. I have created this field to facilitate automatic updates from GenBank records. Most SEA-PHAGES genomes are expected to be automatically updated from GenBank once they are assigned a unique GenBank accession. However, some genomes, such as those

generated from non-SEA-PHAGES researchers, may not need to be automatically updated. This field is set to *1* for genomes that are to be automatically updated and set to *0* for those genomes that are not to be automatically updated.

AnnotationAuthor. I have created this field to indicate whether or not the genome has been derived from the SEA-PHAGES program, and it facilitates automatic updates from GenBank. If a genome has been sequenced or annotated through the SEA-PHAGES program, “Graham Hatfull” is expected to be a listed author. This authorship indicates whether the SEA-PHAGES annotators have authority to update genome annotations in the GenBank record. For SEA-PHAGES genomes, this field is set to *1* (“Graham Hatfull” is an author); otherwise it is set to *0* (“Graham Hatfull” is not an author).

Status. Originally, this field indicated whether a SEA-PHAGES genome had been automatically (*draft*) or manually (*final*) annotated. However, the incorporation of non-SEA-PHAGES genomes into the *Actino_Draft* databases complicates this field, as the strategy of gene annotations by other research groups is not always obvious. I have added a *gbk* status that indicates the annotation status is not known. This field now indicates annotation status for any genome regardless of whether it is a SEA-PHAGES genome or not. The *AnnotationAuthor* field now reflects whether a genome is derived from the SEA-PHAGES program or not.

AnnotationQC. I have created this field to facilitate downstream analyses of annotation data. Since PhameratorDB stores annotations of varying degrees of quality, this field reflects a simple, binary metric of confidence in the annotation data, so that tools (such as Starterator) can rely more heavily on some types of annotations than others. This field is set to *1* for reliable annotations and set to *0* for unreliable annotations.

I have created and modified the following fields in the *Gene* table.

LocusTag. I have created this field to facilitate automatic updating of GenBank records. Once a genome has been submitted to GenBank, genes are assigned unique locus tags in the LOCUS_TAG field. These identifiers cannot be changed, and annotators are required to use them when requesting to update details about individual genes. This field provides a direct link to the corresponding GenBank feature.

I have created the following field in the *Version* table.

Schema. This field enhances version control of scripts that directly communicate with PhameratorDB. As the structure of the database changes, such as by the addition or removal of tables or fields, the database schema number can be incremented to reflect that changes have been made.

In addition to improving the structure of PhameratorDB, I have developed several interactive scripts and tools, as well as improved upon scripts previously developed by Charles Bowman or Steve Cresawn, to improve the database management process (Table A-1). These scripts are written in Python 2.7 and rely on several third-party Python packages or stand-alone command line tools. They should be executed from the command line in Ubuntu OS, and they directly communicate with PhagesDB, GenBank, PECAAN, a local copy of the NCBI CDD, and a local copy of PhameratorDB in MySQL. The complete collection of scripts is stored in a `k_phamerate` repository and is tracked using the version control system, `git` (<https://git-scm.com>). I made the repository publicly available on the SEA-PHAGES GitHub page (https://github.com/SEA-PHAGES/k_phamerate).

Table A-1. Python scripts to maintain PhameratorDB.

Script	Description
<code>retrieve_database_updates.py</code>	Retrieve new data from PhagesDB and GenBank to be imported
<code>import_phage.py</code>	Import new or updated phage genomes and annotations
<code>cdd_pp.py</code>	Identify conserved domains from NCBI CDD for phage genes
<code>k_phamerate.py</code>	Group phage genes into phams based on amino acid sequence similarity
<code>export_database.py</code>	Export PhameratorDB instance and upload to server
<code>freeze_database.py</code>	Create PhameratorDB instance that does not contain any draft genome annotations
<code>update_[field].py</code>	Sub-collection of scripts to update specific fields as needed
<code>compare_databases.py</code>	Directly compare phage data in PhagesDB, a PhameratorDB instance, and GenBank

The **k_phamerate** scripts are designed to be executed separately to enhance flexibility of PhameratorDB management. In general, the **k_phamerate** repository can be used to manage different PhameratorDB instances. However, some scripts, or particular functions within scripts, are only applicable to the primary actinobacteriophage PhameratorDB instance, *Actino_Draft*. Many of the scripts are executed during each round of updates to the PhameratorDB instance. Below is a description of how each script or tool is used during a typical round of database updates to highlight how they function and work together.

A.2.4 Development of ticket tracking systems

With new genome sequences and annotations routinely becoming available for PhameratorDB, and with thousands of researchers routinely using Phamerator, many updates and issues relating to the general management of PhameratorDB, or specifically to the *Actino_Draft* instance, arise. Together, these diverse actions constitute database “tickets”, such as:

- New genome sequences need to be auto-annotated and imported
- New *preliminary final* annotations need to be reviewed and imported
- Updated SEA-PHAGES genomes in GenBank need to be imported
- New non-SEA-PHAGES genomes in GenBank need to be imported
- Changes need to be made to genome information (e.g. host, cluster, etc.) or to gene information (e.g. coordinates, descriptions, etc.)
- Genes need to be added or removed

Systematically maintaining and addressing tickets provides a framework for efficient database management. As a result, I have created two separate ticket tracking systems.

In order to handle and control requests and notifications regarding new database tickets from various sources, I have developed an email-based ticket tracking system with the assistance of Christian Gauthier (Figure A-1). The email account, phamerator.qc@gmail.com, provides a standardized communication channel between the database administrator and end-users to improve database management by tracking, addressing, and storing unstructured tickets in the form of emails. Tickets regarding a) the availability of new genome sequences or new *preliminary final* annotations are submitted from PhagesDB, b) revisions to *final* annotations are submitted from any of the annotators, and c) specific, miscellaneous issues or questions about the

database are submitted from any end-user. All tickets that need to be addressed are stored in the inbox. After they are addressed, the originator (or submitter) of the ticket can be directly notified that it has been addressed. The email account can be used to maintain a history of tickets by systematically storing emails after they are addressed.

I have also developed a second, structured, database ticketing system that is used to import new data, make updates to the database, and maintain a record of updates. This ticketing system plays an integral component to the automated PhameratorDB management system and represents one of the most important innovations I have contributed to the collection of **k_phamerate** scripts (Table A-2). For most types of updates to the database, there needs to be a unique ticket in a csv-formatted **import_table** that provides instructions on how to implement the update. Within the **import_table**, an individual row of data populating 11 columns constructs a unique ticket. There are currently four types of tickets that the import script can implement (*add*, *remove*, *replace*, and *update*), and this information is indicated in Column 1. *Add* indicates that a new phage genome (*Phage #1*) needs to be added to the database. *Remove* indicates that a phage genome (*Phage #2*) currently in the database needs to be removed. *Replace* indicates that a phage genome (*Phage #2*) currently in the database needs to be replaced with a new phage genome (*Phage #1*). *Update* indicates that no genome is being added, removed, or replaced, but that new metadata for a phage genome (*Phage #1*) currently in the database needs to be updated.

The remaining 10 fields provide the necessary information for the ticket to be implemented. Column 2 indicates the name of the new phage genome that *update*, *add*, and *replace* tickets address, and columns 3 through 9 contain metadata pertaining to that phage that will be used to populate fields in the database. Column 6 indicates the stage of gene annotations

(*draft*, *final*, *gbk*). Column 7 indicates whether SEA-PHAGES has control of the annotations (*hatfull* or *gbk*, which get converted to a binary number in the database). Column 8 indicates which field of the CDS features in the associated GenBank-formatted flat file is expected to contain the gene descriptions (*note*, *function*, *product*). Column 10 indicates the run mode (*pecaan*, *phagesdb*, *ncbi_auto*, *ncbi_misc*, *other*, *custom*) for *add* and *replace* tickets, determining which QC checks are used to evaluate the associated flat file. Column 11 indicates the name of the phage genome that will be removed in *replace* and *remove* tickets. Import tickets are automatically generated by the `retrieve_database_updates.py` script, but they can also be manually generated. Details about different types of tickets are described below, and examples of how the data retrieval script creates tickets are provided in Table A-2.

Table A-2. Structure of import tickets.

Ticket column		Example tickets based on data type				
#	Name	Metadata update	New auto-ann.	New prelim. final ann.	Updated GenBank ann.	New non-SEA-PHAGES ann.
1	<i>Ticket type</i>	update	add	replace	replace	add
2	<i>Phage #1</i>	[PhagesDB]	[PhagesDB]	[PhagesDB]	[PhamDB]	[Manual]
3	<i>Host genus</i>	[PhagesDB]	[PhagesDB]	[PhagesDB]	[PhamDB]	[Manual]
4	<i>Cluster</i>	[PhagesDB]	[PhagesDB]	[PhagesDB]	[PhamDB]	UNK
5	<i>Subcluster</i>	[PhagesDB]	[PhagesDB]	[PhagesDB]	[PhamDB]	UNK
6	<i>Annotation status</i>	[PhamDB]	draft	final	[PhamDB, or <i>final</i> if replacing a <i>draft</i>]	gbk
7	<i>Annotation author</i>	[PhamDB]	hatfull	hatfull	[PhamDB]	gbk
8	<i>Gene descriptions field</i>	none	product	product	product	product
9	<i>Accession</i>	[PhagesDB]	none	[PhagesDB]	[PhamDB]	[Manual]
10	<i>Run mode</i>	none	pecaan	phagesdb	ncbi_auto	ncbi_misc
11	<i>Phage #2</i>	none	none	[PhamDB]	[PhamDB]	none

Note: in the table above, bracketed text indicates the source of data used to populate the specified field. Many fields can be populated by the script using data from PhagesDB or PhameratorDB (“PhamDB”), while some fields may need to be manually populated. If *Cluster* and *Subcluster* have not been determined, they can be designated as UNK (“unknown”).

A.2.5 Retrieve new data to import into PhameratorDB

The first step to update PhameratorDB requires gathering all new genome data (new sequences or annotations), creating corresponding tickets, and systematically staging them in a local directory in preparation for evaluation and import into PhameratorDB. This is accomplished using the `retrieve_database_updates.py` script, which requires two arguments (Table A-3).

Table A-3. Required arguments for `retrieve_database_updates.py` script.

Argument	Description
1	Name of database for which to retrieve updates
2	Path to directory where updates are staged for import

This script retrieves four types of data to be imported and creates tickets for them. The database administrator can select to retrieve all, or only specific, types of data. For each type of data, the script stores the retrieved data in a structured directory ready for import: a new folder is created that contains a) one csv-formatted `import_table` listing each ticket, and b) a `genome` subdirectory containing flat files.

Metadata updates. Genome data for all sequenced phages on PhagesDB is retrieved at: http://phagesdb.org/api/sequenced_phages/. For phage metadata, the script iterates through every phage in PhameratorDB, matches the *PhageID* to the identical phage name in PhagesDB, and compares metadata stored in PhameratorDB to what is stored in PhagesDB, including: *Cluster*, *Subcluster*, *Host genus*, and *Accession*. PhagesDB is the primary data source for these fields, so if any information is different between the two databases, a new import ticket is created with the

current metadata from PhagesDB, so that PhameratorDB is synchronized accordingly (Table A-2, *Metadata update*).

New auto-annotations. When new phage genomes are sequenced, the sequence is uploaded to PhagesDB. The *draft* auto-annotations are imported into the *Actino_Draft* instance for immediate reference to end-users (with tools such as the Phamerator GUI and Starterator) until the refined, reliable, *final* annotations are prepared. PhagesDB tracks which genomes have been imported into *Actino_Draft*, and it provides a list of newly sequenced, “unphamerated” phages that need to be imported at: <http://phagesdb.org/data/unphameratedlist>. Automated annotations of these new genomes can be generated through the phage genomics tool, PECAAN (<https://discover.kbrinsgd.org>). For each new phage genome, a request for auto-annotation is sent to PECAAN with the URL: [https://discoverdev.kbrinsgd.org/phameratoroutput/phage/\[PhageID\]](https://discoverdev.kbrinsgd.org/phameratoroutput/phage/[PhageID]) (where [PhageID] indicates the specific phage name of interest). PECAAN retrieves the new sequence from PhagesDB and automatically annotates coding sequence (CDS) genes using Glimmer (Delcher et al., 1999) and GeneMark (Borodovsky et al., 2003), and tRNA genes using tRNAscan-SE (Lowe and Eddy, 1997) and ARAGORN (Laslett and Canback, 2004). The script retrieves a GenBank-formatted flat file of the auto-annotations, stores it in the local staging directory, and creates a new import ticket in the `import_table` (Table A-2, *New auto-ann.*). When auto-annotated *draft* genomes are imported into *Actino_Draft*, PhagesDB removes them from the list of “unphamerated” genomes so that they are not re-processed during subsequent rounds of PhameratorDB updates. It is important to note that since the list of “unphamerated” genomes is created by PhagesDB based on data in the *Actino_Draft* PhameratorDB instance, this data retrieval step is not reliable if it is used to update alternative PhameratorDB instances.

New preliminary final annotations. The *draft* gene annotations are eventually replaced in PhameratorDB with manual, *final*, gene annotations. The refined annotations are submitted by senior annotators as *preliminary final* annotations to PhagesDB in GenBank-formatted flat files so that they can be evaluated for quality in the PhameratorDB pipeline. When *preliminary final* annotations are uploaded to PhagesDB, the flat files are stored on Phagesdb in the *qced_genbank_file* field with a timestamp stored in the *qced_genbank_file_date* field. Similar to new metadata retrieval, the retrieval script iterates through every *PhageID* in PhameratorDB, matches it to the phage name in PhagesDB, reviews the date (if any) that a *preliminary final* annotation file was uploaded, and if it is more recent than the date of the annotations stored in PhameratorDB (indicated in the *DateLastModified* field of the *Phage* table), it retrieves the new flat file from the *qced_genbank_file* field, stages it in the **genome** folder, and creates a new import ticket (Table A-2, *New prelim. final ann.*). The quality of the gene annotations is reviewed during import using the **import_phage.py** script (see below).

Updated GenBank annotations. After *preliminary final* annotations are evaluated for quality and imported into PhameratorDB, they are eventually submitted to GenBank with a unique accession number. Annotators provide the accession number to PhagesDB, which gets retrieved for import into PhameratorDB using the *metadata* retrieval step (see above). In the *update GenBank data retrieval* step, the script iterates through every *PhageID* in PhameratorDB, retrieves the accession number from the *Accession* field (if any), and retrieves the associated genome record from the GenBank *nucleotide* database using the Entrez python package. For each record, the script identifies the date of the record (Figure A-2), and if the record date is more recent than the date of the annotations stored in PhameratorDB (indicated in the *DateLastModified* field of the *Phage* table), the script retrieves the updated flat file, stages it in

the **genome** folder, and creates a new import ticket (Table A-2, *Updated GenBank ann.*). The database administrator is required to provide an email address for this step since the NCBI requests contact information for all users downloading data from GenBank through Entrez. Additionally, the script creates a csv-formatted **summary_table** of all *PhageIDs*, their accession, and the results of data retrieval from GenBank.

New non-SEA-PHAGES annotations. Actinobacteriophage genomes that have been sequenced and annotated outside of the SEA-PHAGES program occasionally become available in GenBank. SEA-PHAGES annotators review the quality of these genomes and annotations and assess whether or not they should be imported into the *Actino_Draft* database. If the genomes should be imported, the database administrator manually retrieves the flat files from GenBank, stages them in a local directory, and creates the appropriate import tickets (Table A-2, *New non-SEA-PHAGES ann.*).

Miscellaneous notes about data retrieval. Currently, the default value populating the *Gene description field* field is **PRODUCT**, since this is where descriptions generated from the SEA-PHAGES program are expected to be stored. Also, if GenBank accession numbers are retrieved from PhagesDB, they are retrieved from the *genbank_accession* field, ensuring that GenBank **ACCESSION** numbers (and not RefSeq **ACCESSION** numbers) are stored in PhameratorDB (see above). Last, it is important to note that not all fields are applicable for every ticket type, but the import script requires that all fields are populated. Fields that do not contain relevant data for the ticket type should be populated with *none* (Table A-2).

A.2.6 Import new data into PhameratorDB

After all types of data are retrieved, it is ready to be evaluated and imported into PhameratorDB using `import_phage.py`. This script was initially created by Charles Bowman to parse data from flat files to add new genomes to PhameratorDB, and I have continued to improve and expand its functionality so that it is now a versatile, invaluable tool for specifically updating the *Actino_Draft* database, managing different PhameratorDB instances, and supporting the entire SEA-PHAGES pipeline. First, the script performs a variety of administrative actions on the database in addition to parsing flat files. Second, it relies on import tickets (such as those generated from the *data retrieval* step) so that the import process is substantially automated. Third, it evaluates incoming data, either from parsed flat files or import tickets, with robust quality control (QC) checks that vary depending on the data type, and only proceeds with adding new data to the database once it passes QC. Fourth, it evaluates and imports data in an interactive environment that provides the database administrator with control over the process. Three arguments are required to run this script (Table A-4).

Table A-4. Required arguments for `import_phage.py` script.

Argument	Description
1	Name of database that will be updated
2	Path to directory where genome flat files are staged
3	Path to import table filename

Step 1: Parse and validate import table. The first step of the import script is to parse and prepare tickets from `import_table`. The script confirms the table is structured appropriately and that the fields are populated correctly for each ticket type. For each ticket type, there are specific rules regarding how the ticket fields are populated to ensure that the ticket is

implemented correctly. For instance, if a new genome is being added with an *add* ticket, the *Phage #2* field should be populated with *none*; if the *add* ticket contains a phage name in this field, an error is encountered. Additionally, the script confirms that there are no duplicated tickets or tickets with conflicting data (such as an *add* and *remove* ticket for the same phage). After import tickets are parsed and evaluated, they are grouped by ticket type and are ready to be implemented.

The rules governing how import tickets are structured ensure that the tickets are implemented appropriately, but they make it more difficult to construct tickets. Import tickets are automatically generated by the `retrieve_database_updates.py` script, and it is recommended to rely on this script as much as possible to minimize potential errors. However, import tickets can still be generated manually, and sometimes it may even be necessary (e.g. adding non-SEA-PHAGES genomes from GenBank). Manual creation of tickets for phages that are in PhagesDB can be slightly simplified. The *HostStrain*, *Cluster*, *Subcluster*, and *Accession* fields can be populated with *retrieve*. When the script encounters this, it retrieves the requisite data for that specific field directly from PhagesDB, using [http://phagesdb.org/api/phages/\[PhageID\]](http://phagesdb.org/api/phages/[PhageID]) (in which [PhageID] refers to the specific *PhageID* from the import ticket).

Step 2: implement *update* tickets. Update tickets are implemented first. The script accesses the PhameratorDB instance indicated in the first script argument, and for each ticket, it updates data in *HostStrain*, *Status*, *Accession*, *AnnotationAuthor*, *Cluster2*, *Subcluster2*, and *Cluster* in the *Phage* table for the *PhageID* that matches *Phage #1* in the ticket.

Step 3: implement *remove* tickets. Remove tickets are implemented next. For each *remove* ticket, all data (predominantly in the *Phage*, *Gene*, and *Pham* tables) associated with *Phage #2* are deleted from the database.

Step 4: implement *add* and *replace* tickets. The largest step in the import script involves implementing these two types of tickets. In this step, the script retrieves all files stored in the **genome** folder indicated by the second script argument, confirms they have a permissible file extension (**gb**, **gbf**, **gbk**, **txt**), and confirms that they are a GenBank-formatted flat file that can be parsed (Figure A-2). When new annotations become available for a genome, instead of identifying the exact changes and implementing them in PhameratorDB, all data pertaining to that genome is completely replaced using data from the most up-to-date flat file. This process is slightly more complex than simply deleting the old data and adding the new data, as there are QC checks to ensure that the old genome indicated in the ticket (*Phage #2*) is indeed the correct genome in the database to be replaced by the new genome (*Phage #1*). Two types of data are parsed from the flat file and evaluated (Figure A-2).

The first type of data relates to the entire phage genome, such as the phage name, nucleotide sequence, host genus, accession, and authorship. The data in the flat file is matched to the import ticket by the phage name parsed from the SOURCE-ORGANISM field at the top of the file, and subsequently evaluated and compared to data in the import ticket and in PhameratorDB. After this, several fields in the *Phage* table are populated from data derived from the import ticket, from the flat file, or from the script itself (Table A-5).

Table A-5. Source of data used to populate *Phage* table.

Table Field	Data Origin	Description
<i>PhageID</i>	Ticket	Unique name of phage in database
<i>Name</i>	Ticket	Name of phage (with <i>_Draft</i> suffix if <i>draft</i> status)
<i>HostStrain</i>	Ticket	The host genus
<i>Cluster</i>	Ticket	The assigned subcluster OR cluster
<i>Cluster2</i>	Ticket	The assigned cluster
<i>Subcluster2</i>	Ticket	The assigned subcluster
<i>Accession</i>	Ticket	The accession indicating origin of annotation data or used for automatic updates from GenBank
<i>Sequence</i>	Flat file	The genome nucleotide sequence
<i>Sequence Length</i>	Flat file	Computed length of the nucleotide sequence
<i>Sequence GC%</i>	Flat file	Computed GC% content of the nucleotide sequence
<i>DateLastModified</i>	Script	The date that the genome annotations are imported into the database
<i>Status</i>	Ticket	Indicates the annotation status of the genome record
<i>AnnotationAuthor</i>	Ticket	Indicates whether “Graham Hatfull” is a listed author
<i>RetrieveRecord</i>	Ticket or Script	Indicates whether the record should be automatically updated from GenBank records
<i>AnnotationQC</i>	Ticket or Script	Indicates reliability of annotations for downstream applications

Matching tickets to flat files requires that the phage names are spelled identically. Sometimes this is not the case, in which the desired spelling of the phage name in PhameratorDB (and thus in the import ticket) is slightly differently than the spelling in the GenBank record. These conflicts can arise for several reasons that cannot be immediately corrected. This includes slight variations in nomenclature (such as “phiELB20” versus “ELB20”), inadvertent typos introduced (such as “Fionnbarth” versus “Fionnbharth”), different nomenclature constraints implemented in GenBank (such as “LeBron”, which is spelled “Bron” in the SOURCE-ORGANISM field of the GenBank record), or different nomenclature constraints implemented in PhameratorDB or PhagesDB (such as “ATCC29399B_C” versus “ATCC29399BC”). To account for these conflicts, the import script contains a hard-coded phage name dictionary that converts several GenBank phage names to the desired phage name stored in PhagesDB and the *Actino_Draft* database. This list contains fewer than 10 name conversions and does not change frequently. I have also developed an alternative strategy that circumvents this issue, in which the

phage name is parsed from the filename of the flat file instead of from the SOURCE-ORGANISM field within the record. This allows for greater flexibility when parsing batches of flat files that may not adhere to these default expectations, such as when new PhameratorDB instances are developed for phages that have not been annotated through the SEA-PHAGES program. This option can be implemented using a different run mode (which is discussed in greater detail in the *Run Mode* section below).

The *AnnotationQC* field provides a binary metric of the reliability of the gene annotations for downstream applications. Initially, this field is indirectly determined by the *Annotation Status*. For genomes being replaced or newly added, if the genome's *Annotation Status* = *final*, *AnnotationQC* is set to *1*, otherwise it is set to *0*. However, if this field is manually changed outside of the normal update pipeline (such as with the `update_[field].py` scripts discussed below), the new value is retained during this step.

The *RetrieveRecord* field provides a binary indicator of whether the genome should be automatically updated from GenBank. Initially, this field is indirectly determined by the *AnnotationAuthor* field. For newly added genomes, if *AnnotationAuthor* = *hatfull*, this field is set to *1*, otherwise it is set to *0*. For genomes being replaced (by automatic updates from GenBank or by the creation of manual tickets), the value in this field is retained.

Several fields in the flat file contain data about the phage and host names: DEFINITION, SOURCE, SOURCE-ORGANISM, and the ORGANISM, HOST, and LAB_HOST sub-fields of the FEATURE-SOURCE field (Figure A-2). The host and phage name data stored in PhameratorDB is derived from the ticket, but they are compared to the data parsed from these various fields for confirmation.

The second type of data parsed from the flat file pertains to individual genes (and is stored in the *Gene* table). After parsing the phage genome information, the script iterates through the annotation features in the file. The tRNA and CDS features are evaluated, and all others are ignored.

Currently, tRNA features are not stored in PhameratorDB. However, it is valuable to evaluate the basic characteristics of these features in *preliminary final* annotations prior to submission to GenBank. The tRNAs are expected to be between 60 and 100 nucleotides long, should have a terminal A or C nucleotide, and should have a PRODUCT field with a codon indicated. The script issues a warning if these restrictions are not met. More robust tRNA QC checks can be developed if these features eventually need to be stored in PhameratorDB.

CDS features are parsed and stored in the *Gene* table, and the script utilizes many QC checks to confirm their integrity. The majority of data that the import script stores in the *Gene* table are derived directly from the flat file (Table A-6).

Table A-6. Source of data used to populate *Gene* table.

Table Field	GenBank CDS Feature Field	Description
<i>GeneID</i>	LOCUS_TAG or <i>custom</i>	Unique name of CDS in database
<i>PhageID</i>	N/A	<i>PhageID</i> in <i>Phage</i> table
<i>Start</i>	LOCATION	Start coordinate of the feature
<i>Stop</i>	LOCATION	Stop coordinate of the feature
<i>Length</i>	LOCATION	Computed size of nucleotide feature, determined from amino acid sequence
<i>Name</i>	LOCUS_TAG or <i>custom</i>	CDS number
<i>TypeID</i>	FEATURE	The database retains data only for CDS features
<i>Translation</i>	TRANSLATION	Amino acid sequence
<i>Orientation</i>	STRAND	Strand orientation of feature (<i>Forward</i> , <i>Reverse</i>)
<i>Notes</i>	PRODUCT, FUNCTION, or NOTE	The field containing the gene descriptions
<i>LocusTag</i>	LOCUS_TAG	The official name for the feature in the GenBank record. This is not retained for annotations that were not retrieved from GenBank

The *GeneID* is the primary key in the *Gene* table and represents a unique name of the gene in the database. This can be derived three ways. First, it can simply be synonymous with the LOCUS_TAG of the CDS feature in the flat file. For SEA-PHAGES flat files, this is usually the case. However, for non-SEA-PHAGES flat files, there may not be a LOCUS_TAG for every, or any, CDS feature. As a result, the *GeneID* can be computed on-the-fly by concatenating the *PhageID* with the CDS count (which indicates the order that the CDS was parsed from the feature list). The import script uses the ticket's run mode to determine which of these two strategies is implemented. However, neither of these naming strategies guarantee the *GeneID* is unique in the *Gene* table, and naming conflicts may arise with features already present in the *Gene* table. In this case, a warning is issued and a *_duplicateID[0123]* suffix is appended to the *GeneID* (where [0123] is an integer).

Gene descriptions are stored in the *Notes* field of the *Gene* table. However, CDS features in flat files can contain descriptions in three different fields: PRODUCT, FUNCTION, and NOTE. The *Gene description field* field in the import ticket indicates which of these three fields are expected to contain gene description data in the flat file. If the script identifies gene descriptions in the other two fields as it parses CDS features, it issues a warning.

The *LocusTag* field in the *Gene* table is populated directly from the LOCUS_TAG field in the CDS feature. However, unlike the *GeneID* field, the *LocusTag* field does not need to contain unique, non-duplicated values. Storing the LOCUS_TAG data provides an unambiguous link to the original CDS feature in the GenBank record, regardless of the restrictions imposed on the CDS feature's *GeneID*. This is valuable when reporting the gene information in a publication, and it is required when requesting GenBank to update information about specific CDS features (such as corrections to coordinates or gene descriptions).

In many GenBank records, CDS features may contain descriptions that are not informative, including “hypothetical protein”, “phage protein”, “unknown”, “conserved hypothetical protein”, ordered numerical data, “gp[0123]” and “ORF[0123]” (where [0123] is an integer), and “putative protein”. These generic descriptions are not retained in PhameratorDB.

Many QC steps in the import script need to be performed on every genome (such as confirming the nucleotide sequence is not already present in the database under a separate name). However, since the *Actino_Draft* database stores data for diverse types of genomes, and some QC steps are dependent on factors such as the annotation status (*draft*, *final*, *gbk*), the authorship (*hatfull* or *non-hatfull*), or the data source (such as PhagesDB or GenBank). As a result, I have created 11 QC steps that can be toggled on (*yes*) and off (*no*) depending on the type of genome being imported.

use_basename. By default, phage names in flat files are expected to be in the SOURCE-ORGANISM field. When this QC option is selected, the name of the file (without the file extension) is used as the phage name (*yes* = filename is used). This option is useful when importing non-SEA-PHAGES genomes.

custom_gene_id. By default, the *GeneID* is derived from the LOCUS_TAG. When this QC option is selected, the *GeneID* is created by concatenating the *PhageID* and CDS count (*yes* = the *GeneID* is created by concatenation). This option is useful when importing non-SEA-PHAGES genomes.

ignore_gene_id_typo. By default, a warning is issued if a *GeneID* does not contain the phage name, indicating there is likely a typo in the *GeneID*. When this QC option is selected, this warning is silenced (*yes* = *GeneID* spelling is ignored). This option is useful when importing

genomes from GenBank; since the GenBank LOCUS_TAG cannot be changed, there is no need for the script to issue warnings.

ignore_description_field_check. By default, a warning is issued if gene descriptions appear to be present in fields other than the field indicated by the import ticket. When this QC option is selected, this warning is silenced (*yes* = import gene description data from the indicated ticket field without checking other fields). This option is useful when importing SEA-PHAGES genomes from GenBank, which have been systematically annotated with descriptions in the PRODUCT field.

ignore_replace_warning. By default, a warning is issued if a genome with *final* status is about to be replaced with a new genome. When this QC option is selected, this warning is silenced (*yes* = *final* status is ignored). This option is useful when importing genomes from GenBank, when it is expected that *final* status genomes will be replaced.

ignore_trna_check. By default, tRNA features are evaluated for quality, and warnings are issued when problems are encountered. When this QC option is selected, these warnings are silenced (*yes* = tRNA QC is ignored). This option is useful when importing *draft* status genomes or genomes from GenBank.

ignore_locus_tag_import. By default, data from the GenBank LOCUS_TAG field is stored in the *Gene* table *LocusTag* field. However, the *LocusTag* field should only reflect data from official GenBank records. When this option is selected, LOCUS_TAG data is not imported (*yes* = locus tags are ignored). This option is useful when importing any genome that has not been obtained from GenBank.

ignore_phage_name_typos. By default, a warning is issued if any of the various phage name fields in the flat file contain phage name typos. When this option is selected, the warning is

silenced (*yes* = phage name typos are ignored). This option is useful when importing non-SEA-PHAGES genomes from GenBank.

ignore_host_typos. By default, a warning is issued if any of the various host name fields in the flat file contain host name typos. When this option is selected, the warning is silenced (*yes* = host genus typos are ignored). This option is useful when importing non-SEA-PHAGES genomes from GenBank.

ignore_generic_author. By default, a warning is issued if the author field in the flat file contains a generic author “Lastname, Firstname”, which can be inadvertently added during genome annotation. When this option is selected, the warning is silenced (*yes* = generic authors are ignored). This option is useful when importing *draft* status genomes, or genomes from GenBank.

ignore_description_check. By default, a warning is issued if gene descriptions appear to contain errors (although, this QC step is currently under-developed). When this option is selected, the warning is silenced (*yes* = gene description errors are ignored). This option is useful when importing *draft* status genomes or genomes from GenBank.

Although there are currently 11 optional QC steps, more may be added as the database grows in complexity. In order to manage which optional QC steps are implemented, I created run modes that are specified for each ticket (Table A-7).

Table A-7. Quality control options differ between run modes.

QC Option	pecaan	phagesdb, other	ncbi_auto	ncbi_misc
use_basename	no	no	no	yes
custom_gene_id	no	no	no	yes
ignore_gene_id_typo	no	no	yes	yes
ignore_description_field_check	no	no	yes	no
ignore_replace_warning	no	no	yes	yes
ignore_trna_check	yes	no	yes	yes
ignore_locus_tag_import	yes	yes	no	no
ignore_phage_name_typos	yes	no	yes	yes
ignore_host_typos	yes	no	no	yes
ignore_generic_author	yes	no	yes	yes
ignore_description_check	yes	no	yes	yes

The *pecaan* run mode is used for *draft* annotations. The *phagesdb* run mode is used for SEA-PHAGES *preliminary final* annotations retrieved from PhagesDB. The *ncbi_auto* run mode is used for SEA-PHAGES *final* annotations retrieved from GenBank. The *ncbi_misc* run mode is used for non-SEA-PHAGES annotations retrieved from GenBank. The *other* run mode is reserved for when database administrators are not sure which run mode is appropriate; it currently defaults to the *phagesdb* run mode. Lastly, the *custom* run mode enables the database administrator to manually select which of the 11 QC steps should be performed if none of the other four preset run modes are appropriate. As the database grows in complexity additional run modes may need to be created.

As QC steps are performed on tickets, the genome either passes or fails QC. When some QC steps are not met, an error is issued. In contrast, when some QC steps are not met, the script pauses and issues a warning, requiring the administrator to indicate whether an error should be issued or not. If a genome acquires one or more errors during import, the entire genome fails to be imported, and no changes are made to the database for that genome. The success or failure of an import ticket has no impact on the success or failure of the next ticket, and the script iterates

through all *add* and *replace* tickets. After all *add* and *replace* tickets are processed, the script is completed.

I have created several methods to tracking and managing tickets (and the associated genomes) as they pass or fail QC. First, a summary of the import process is reported to the user in the UNIX shell during import and after all tickets are processed. Second, the results of every ticket are recorded in a **log** file, including any errors and warnings that were generated. Third, tickets and genome files are moved to new folders based on their import status. All tickets that were successfully implemented with no errors are recorded in a **successful_import_table**, and the associated genomes are moved to a **successful_genomes** folder. In contrast, all tickets that failed QC due to one or more errors are recorded in a **failed_import_table**, and the associated genomes are moved to a **failed_genomes** folder. This enables quick reference to the specific tickets and genome files that need to be reviewed, modified, and repeated. Fourth, I have created *test* and *production* run types that the administrator can choose between. During a *production* run, import tickets and genome files are processed and evaluated, and the database is updated as specified by the ticket if QC is passed. In contrast, during a *test* run, import tickets and genome files are processed and evaluated, but the database is not updated. The *test* run is a valuable tool to determine whether any particular group of tickets and flat files are ready to be imported without actually altering the database. The import script can be executed on the same tickets and flat files multiple times, each time making the appropriate modifications until the ticket contains no errors. Many SEA-PHAGES annotators now rely on the import script *test* run to personally evaluate *preliminary final* annotations prior to uploading them to PhagesDB, and this has helped to improve the speed and efficiency of the PhameratorDB pipeline.

The import script is designed to handle diverse types of tickets present in a single `import_table`. However, the `retrieve_database_updates.py` script creates separate staged directories and import tables for different types of data to be imported to minimize potential ticket conflicts. When the `import_script.py` is executed following the `retrieve_database_updates.py` script, it is recommended that the script is executed separately for each ticket type, and in the following order: metadata updates, auto-annotated genomes from PECAAN, new *preliminary final* annotations from PhagesDB, auto-updated SEA-PHAGES *final* annotations from GenBank, and other miscellaneous tickets that need to be implemented.

A.2.7 Update specific fields

Although it is recommended that genome and gene data are imported and updated directly from flat files using the import script, sometimes it may be necessary to modify or update specific fields for specific phages. This can be accomplished using a collection of `update_[field].py` scripts, in which the `[field]` refers to the specific field that needs to be updated (Table A-8). Charles Bowman created the first update script, and I improved upon it and created others that constitute the current `update_[field]` collection.

Table A-8. Structure of update_[field] tables.

Script	Table	Column 1	Column 2
update_phageid.py	<i>Phage</i>	<i>PhageID</i>	<i>PhageID</i>
update_phagename.py	<i>Phage</i>	<i>PhageID</i>	<i>Name</i>
update_accession.py	<i>Phage</i>	<i>PhageID</i>	<i>Accession</i>
update_annotation_author.py	<i>Phage</i>	<i>PhageID</i>	<i>AnnotationAuthor</i>
update_ncbi_retrieval_flag.py	<i>Phage</i>	<i>PhageID</i>	<i>RetrieveRecord</i>
update_datelastmodified.py	<i>Phage</i>	<i>PhageID</i>	<i>DateLastModified</i>
update_host.py	<i>Phage</i>	<i>Name</i>	<i>HostStrain</i>
update_status.py	<i>Phage</i>	<i>Name</i>	<i>Status</i>
update_geneid.py	<i>Gene</i>	<i>GeneID</i>	<i>GeneID</i>
update_gene_description.py	<i>Gene</i>	<i>GeneID</i>	<i>Notes</i>

Each script is similarly structured, is pointed to the appropriate table in the database to implement updates, and requires two arguments (Table A-9). In contrast to the 11-field import tickets, a list of simple update tickets are stored in a csv-formatted **update_table** containing two columns. The first column contains the primary key in the database table for which the target field will be updated, and the second column contains the new data that will populate the target field. For example, in the **update_gene_description.py** script, gene descriptions for hundreds of genes from hundreds of phages can be updated at once; the **update_table** contains a list of unique *GeneIDs* that are derived from the *Gene* table and accompanying gene descriptions, and the script implements these changes in the *Gene* table. Unlike import tables though, the csv-formatted update tables need to be manually constructed.

Table A-9. Required arguments for update_[field].py script.

Argument	Description
1	Name of database in which to update data
2	Path to filename containing update data

It is also important to note that this collection of scripts is under-developed. They do not perform any QC checks on the data in the file. For instance, if a *GeneID* from the `update_table` is not found in the *Gene* table in the database, the ticket fails to be implemented but no warning or error is issued. Therefore, although these scripts are valuable, it is important to use caution when utilizing them.

A.2.8 Identify conserved protein domains

The NCBI maintains a conserved domain database (CDD) in which the functions of protein sequences are systematically categorized and organized (Marchler-Bauer et al., 2011). Every gene product in PhameratorDB is evaluated for conserved domains using a local copy of the CDD (Figure A-1), and this conserved domain data is stored in the database using the `cdd_pp.py` script. This script was created by Charles Bowman, and I have made relatively few changes to it aside from a few required maintenance updates. The script requires two arguments (Table A-10).

Table A-10. Required arguments for `cdd_pp.py` script.

Argument	Description
1	Name of database in which proteins should be tested for conserved domains
2	Path to directory storing the NCBI conserved domain database

In the *Gene* table, there is a field called *Cdd_Status*. When new phage genomes are added to PhameratorDB, the *Cdd_Status* field for each new gene is set to 0. The `cdd_script.py` script retrieves gene products (stored in the *Translation* field of the *Gene* table) for all genes with *cdd_status* < 1. The `rpsblast+` package is used to identify conserved domains using BLAST with an *e*-value threshold = 0.001. For each gene, retrieved CDD data is inserted into the *Domain*

and *Gene_Domain* tables, and the *Cdd_Status* field in the *Gene* table is set to 1 so that this gene is not re-processed during subsequent rounds of updates.

A.2.9 Group all proteins into phamilies

After phage genomes are added to PhameratorDB, the new gene products need to be grouped with the other genes into phamilies (“phams”) (Figure A-1). This is accomplished using the `k_phamrate.py` script (not to be confused with the name of the entire script repository, `k_phamrate`), which requires only one argument (Table A-11). This script was created by Charles Bowman and has not been modified by me.

Table A-11. Required arguments for `k_phamrate.py` script.

Argument	Description
1	Name of database in which proteins should be grouped into phamilies

In general, the script groups genes into phamilies using a kmer-based strategy implemented with `kclust` (Hauser et al., 2013). First, it keeps track of genes and phams present in the database prior to a new round of clustering. Using all gene products in the database, it runs `kclust` with sensitivity of ~ 70% identity covering 25% of the two proteins (using the following options: `-s 3.53 -c 0.25`). New phams are inserted into the database. For each pham, gene products are aligned using `kalign`. The alignments are used to create markov models of the phams using `hhmake`, a consensus sequence is created using `hhconsensus` from the HH-suite software package (Remmert et al., 2011), and the output sequences from `hhconsensus` are re-clustered with `kclust` (using the following options: `-s 0 -c 0.5`). The new phams after the

second iteration are inserted into the *Pham* table of the database, and pham colors that reflect unique phams are computed and inserted into the *Pham_Color* table of the database. The script attempts to maintain consistency of pham designations and colors between rounds of clustering.

A.2.10 Export updated database

After all new data is imported into the database, new CDD data is retrieved, and new genes are grouped into phams, the database needs to be uploaded to a public server so that it is accessible by other tools and applications (Figure A-1). In order to accomplish this, I created the `export_database.py` script, which requires four arguments (Table A-12). There are several steps in this script, and each can be performed separately.

Table A-12. Required arguments for `export_database.py` script.

Argument	Description
1	Name of database to be exported
2	Path to the <code>current</code> directory where new database file will be stored
3	Path to the <code>backup</code> directory where new database file will be stored
4	Path to the <code>output</code> directory where database query information will be stored

Step 1: export the database. The updated PhameratorDB database is stored within the MySQL directory on the database administrator's local computer. In order to provide the updated database to other users, the database needs to be exported from MySQL into a single file that can be easily uploaded to a server (e.g. `Actino_Draft.sql`). In PhameratorDB, the database version is tracked as an integer in the *Version* field of the *Version* table. During export, the script can also update the version (in which the integer is incremented by 1) and create a version file (e.g. `Actino_Draft.version`). This text file contains an integer that corresponds to the database

version integer. The database and version files are stored in the **current** directory indicated in the second script argument. Downstream applications rely on the name of the database to remain unchanged, so filenames in this directory do not change. Since the newly-exported files represent the most up-to-date, current version of the PhameratorDB instance, every round of exports overwrites the two files in this folder. As a result, a backup copy of the new database is also stored in the **backup** directory indicated in the third script argument, but the filename is modified to indicate the version (e.g. **Actino_Draft_v[0123].sql**, where [0123] is an integer). These backup copies can be re-imported into MySQL if needed to return to an older version of the database.

Step 2: Query the new database for updated genome and gene data. Managing PhameratorDB often requires quick reference to data in the *Phage* and *Gene* tables. This step queries the new database and outputs commonly-referenced fields from the *Phage* table (such as *Cluster2*, *Subcluster2*, *HostStrain*, *DateLastModified*, etc.) into a csv-formatted table with a filename that contains the date of retrieval, the database name, and the database version (e.g. **20180101_Actino_Draft_v[0123]_genomes.csv**, where [0123] is an integer). It also outputs commonly-referenced fields from the *Gene* table (such as *GeneID*, *Gene Description*, etc.) into a csv-formatted table with a similarly structured filename (e.g. **20180101_Actino_Draft_v[0123]_genes.csv**, where [0123] is an integer). The files are stored in the **output** directory indicated in the fourth script argument. This step is merely for convenience and does not impact any other process of database administration.

Step 3: Upload database to the server. Once the database and version files are created, they can be uploaded to the Hatfull lab's public server on WebFaction (http://phamerator.webfactional.com/databases_Hatfull). The script uploads the two files

associated with the database that is indicated in the first script argument from the **current** directory that is indicated in the second script argument.

Each step in this script is independent of the others. For instance, data can be queried from a particular database without exporting it or incrementing the version number, and a database file that has been previously exported can be uploaded to the server without having to export it a second time.

A.2.11 Managing the record of tickets

The successful import tickets from the import script and the update tickets from any of the update scripts represent a record of the diverse types of changes made to the database during the round of updates. All tickets and associated files used in each round of updates can be stored in an **updates_history** folder in the same directory as the **current** and **backup** database folders. Direct changes to the database can be made in MySQL software without using scripts in the **k_phamerate** repository, but since this does not generate a record of the changes, it is not recommended as common practice. However, descriptions of these types of changes can also be manually recorded in text files and stored in the **updates_history** folder as well.

A.2.12 Freeze database for publication

The primary PhameratorDB instance, *Actino_Draft*, contains all available actinobacteriophage data, including *draft* and *final* annotations, and is routinely updated and modified. However, research projects may require a version of the database that contains no *draft* genomes, that is no longer routinely modified/updated, and that has a unique identifier from

other database versions or instances. In order to create these “frozen” databases, I created the `freeze_database.py` script, which requires two arguments (Table A-13).

Table A-13. Required arguments for freeze_database.py script.

Argument	Description
1	Name of database to be frozen for publication
2	Path to directory where new database folders will be created

The script copies the database indicated in the first script argument, deletes all data pertaining to *draft* genomes (and thus retains all *final* and *gbk* genomes), and saves the new database with a unique identifier that indicates both the type of phage database and the number of genomes (e.g. *Actinobacteriophage_[0123]*, where *[0123]* is an integer). The script creates a folder for this new database in the directory indicated by the second script argument, and creates three subfolders (**current**, **backup**, and **updates_history**) analogous to the folders regularly used to maintain the *Actino_Draft* instance. Since the new frozen database no longer contains *draft* genomes, gene phams need to be recomputed. As a result, the script does not export the new database, and it resets the database version to 0. The `k_phamerate.py` script can now be executed on this new database, and the `export_database.py` script can be used to increment the version to 1 and upload it to the server. Since it has a unique name, it will not overwrite other databases stored in the local computer or on the server.

Different types of databases may need to be frozen. For instance, sometimes all actinobacteriophages (other than *draft* genomes) need to be retained. Other times, only the *Mycobacterium* phages need to be retained. For my research, I have created databases that contain all phages (regardless of host phylum), or only phages that infect cyanobacterial hosts (Chapter 2). As a result, the script prompts you to choose an appropriate database name (e.g.

Actinobacteriophage, *Bacteriophage*, *Cyanobacteriophage*, etc.) before it appends the genome count to the database name. The database administrator can also input a customized database name if needed. However, this script does not yet provide the functionality of enabling the database administrator to indicate which types of phages should be retained (other than the annotation status), and this step currently needs to be completed with the `mysql` command line utility. It is also important to note that although this database is regarded as “frozen”, it is still able to be modified. Therefore, if minor errors are encountered in the database that need to be modified, the database can be adjusted in MySQL, the version number can be incremented, the database can be re-exported, and the updated database file can overwrite the older file on the server. The script creates the `updates_history` folder to keep track of tickets specific to this PhameratorDB instance similar to the ticket tracking for *Actino_Draft*.

A.2.13 Compare phage databases for consistency

Since three separate databases (PhagesDB, PhameratorDB, and GenBank) store actinobacteriophage genomics data, it is inevitable that data inconsistencies between them will arise unless a mechanism to synchronize all three is established. Although the import script utilizes many QC checks for this purpose, it only ensures consistency for the specific genomes being imported and it only cross-checks databases as directed by the import ticket. In order to ensure comprehensive consistency, I created the `compare_databases.py` script, which requires two arguments (Table A-14).

Table A-14. Required arguments for compare_databases.py script.

Argument	Description
1	Name of database to be checked
2	Path to directory where consistency reports are generated

This object-oriented script enables the database administrator to compare PhameratorDB data to either PhagesDB data or GenBank data, or to compare data between all three databases (it does not compare PhagesDB data to GenBank data unless it also compares PhameratorDB data). All phage data, or subsets of phage data, can be compared. For instance, the administrator can compare genomes depending on their annotation status (*draft*, *final*, or *gbk*) or their authorship (*hatfull*, *non-hatfull*).

Using the PhameratorDB instance indicated by the second script argument, the script retrieves data stored in the *Phage* and *Gene* tables. PhagesDB data for all sequenced phages are retrieved from: http://phagesdb.org/api/sequenced_phages/. All GenBank records are retrieved using accession numbers stored in the *Accession* field of the *Phage* table in PhameratorDB. The script matches all data in PhameratorDB and PhagesDB using the *PhageID* field in PhameratorDB and the phage name in PhagesDB, and it matches all data in PhameratorDB and GenBank using the *Accession* in PhameratorDB. After retrieving and matching data from all databases, the script compares genome data (e.g. phage name, host strain, genome sequence, etc.) and gene data (e.g. locus tags, coordinates, descriptions, etc.), and generates several results files (Table A-15). Additionally, the script can output genomes retrieved from all three databases for future reference and analysis if selected by the user.

Table A-15. Files generated from `compare_databases.py` script.

Output	Description
Accession errors	A list of accession numbers that the script was unable to retrieve GenBank records with (indicating either that an error occurred when annotators recorded the accession number in PhagesDB, that the GenBank record is not yet active, or that there is an error with the GenBank record itself)
Unmatched genomes	A list of PhameratorDB genomes that failed to be matched to PhagesDB or GenBank genomes (perhaps due to non-unique phage names or accession numbers)
Comparison summary	A csv-formatted table summarizing the various types of comparisons performed, and the number of errors encountered for each comparison
Genome errors	A csv-formatted table indicating the specific errors encountered for each PhameratorDB genome
Gene errors	A csv-formatted table indicating the specific errors encountered for each PhameratorDB gene
PhameratorDB genomes	Folder containing all PhameratorDB genomes stored in fasta-formatted files
PhagesDB genomes	Folder containing all PhagesDB genomes stored in fasta-formatted files
GenBank genomes	Folder containing all GenBank records stored in GenBank-formatted flat files

A.3 CONCLUSIONS

The collection of scripts in the `k_phamerate` repository have substantially improved and automated the PhameratorDB data management pipeline. It now provides a faster, more efficient, and more versatile toolkit to manage the progression of a new genome sequence to a GenBank-submitted *final* annotation. It also improves data integrity and accuracy across the entire SEA-PHAGES pipeline by applying robust QC checks between all three phage databases.

Despite these improvements, several steps remain manual and time-intensive. After tickets are implemented, annotators need to be notified whether or not the ticket was implemented successfully. This requires sending an email response for each ticket through the email ticket tracking system, which continues to demand more time as the volume of tickets continues to increase.

It would be valuable for PhameratorDB to retain tRNA data so that they can be evaluated in downstream analyses or visualized in the Phamerator GUI. To accomplish this, PhameratorDB needs to be improved to handle these features, and the `import_phage.py` script needs more robust tRNA QC checks.

The `import_phage.py` script handles *GeneID* duplications by simply appending a `_duplicateID[0123]` suffix (where `[0123]` is an integer). However, this is a cumbersome solution. A more efficient solution is to completely revamp the *GeneID* field, in which all *GeneIDs*, by definition, are a concatenation of the *PhageID* and an integer that increments with each CDS processed. The use of a CDS feature's LOCUS_TAG as the *GeneID* is no longer needed since a *LocusTag* field has been created.

The `compare_databases.py` script has been invaluable in identifying inconsistencies between databases, but it outputs large csv-formatted files for genome and gene results and they can be difficult to interpret. A refined output is needed.

The collection of `update_[field].py` scripts should all be combined into a single `update_field.py` script. This single script would perform the same function as all of the individual scripts but enable the user to select which type of data is being updated, and it would perform more robust QC checks on the data prior to updating the database.

Currently, the use of Phamerator is limited to data derived directly from the SEA-PHAGES program. However, it is an invaluable tool to directly compare the genetic relationships between any types of phages, and with few modifications it could become a tool that is widely available to any phage researcher regardless of whether they are within or without the SEA-PHAGES program.

APPENDIX B OLIGONUCLEOTIDES USED IN THIS STUDY

Table B-1. Oligonucleotides used in this study.

Name	Sequence
oTM13	CATGGAGGGACATATGAGCGGCAAAATC
oTM14	GCACACGGCCAAGCTTGTTTCGG
oTM17	TGAGCCTGACTTGACATCGCACGGCGGGG
oTM18	CCCCGCCGTGCGATGTCAAGTCAGGCTCA
oTM19	TGACCTTGACTTGACATCGCACACTGGGAG
oTM20	CTCCCAAGTGTGCGATGTCAAGTCAAGGTCA
oTM21	TAGTTCCAAGTGTGACACTCATCAGTGAGGG
oTM22	CCCTCACTGATGAGTGTCAAGTTGGAATA
oTM23	TGATCGCAAGTGTGACACCCACCGTGAGGGG
oTM24	CCCTCACGGTGGGTGTCAAGTTGCGATCA
oTM29	GCAGCTACCGGCATCCAGGACACCTGGAAC
oTM30	GTTCCAGGTGTCTGGATGCCGGTAGCTGC
oTM31	TAAGGCGAAGTGTGACATCCACCGGGTTGGG
oTM32	CCCAACCCGGTGGATGTCAAGTTTCGCCTTA
oTM33	TGAGCAGTACTTGACATTCACCAGGTTGGG
oTM34	CCCAACCTGGTGAATGTCAAGTACTGCTCA
oTM35	TGAGCAGTACTTGACATCGCACAGGTTGGG
oTM36	CCCAACCTGTGCGATGTCAAGTACTGCTCA
oTM37	TGAGCAGTACTTGACATCCAACAGGTTGGG
oTM38	CCCAACCTGTTGGATGTCAAGTACTGCTCA
oTM39	TGAGCAGTACTTGACATCCCCCAGGTTGGG
oTM40	CCCAACCTGGGGGATGTCAAGTACTGCTCA
oTM41	TGAGCAGTACTTGACATCGACCAGGTTGGG
oTM42	CCCAACCTGGTCGATGTCAAGTACTGCTCA
oTM43	TGAGCAGTACTTGACATCCACCAGGTTGGG
oTM44	CCCAACCTGGTGGATGTCAAGTACTGCTCA
oTM45	TGAGCAGTACTTGACATCCCACAGGTTGGG
oTM46	CCCAACCTGTGGGATGTCAAGTACTGCTCA
oTM47	TGAGCAGTACTTGACATCGAACAGGTTGGG
oTM48	CCCAACCTGTTTCGATGTCAAGTACTGCTCA
oTM49	TGAGCAGTACTTGACATCGCCCAGGTTGGG
oTM50	CCCAACCTGGGCGATGTCAAGTACTGCTCA
oTM51	GTGGCCAGACGACATCGAGGAGCTTCTCTCGGAGCCCGACTACAAGGAC GACGATGACAAGTAAAAGTCACGACCGGTTGTGTGAGCCAACCCAGGC
oTM52	GTGGCCAGACGACATCGAGGAGCTTCTCTCGGAGCCCTACCCATACGACG TCCCAGACTACGCTTAAAAGTCACGACCGGTTGTGTGAGCCAACCCAGGC
oTM53	CGGGTGAACGAGCACACCAACCTCACCGCCGAGGGTGAACCTCTGGTC GTGGCCAGACGACATCGAGGAGCTT
oTM54	AACAGCTTTGCGAGCCGAGTGAGGGGCACGGGGTTTCCTTTCGTTGCGCG GCCTGGGTTGGCTCACACAACCGGT

oTM55	CAAGAAGAAGCGCCTATTGTCGTGGTG
oTM59	CTTGTCATCGTCGTCCTTGTAAGTC
oTM60	GTAGTCTGGGACGTCGTATGGGTA
oTM69	ACGTAGTGCCCTTTACAGCCACCGAG
oTM70	CCCGCCTGATCTCACCGGTCCAAGTTG
oTM72	GGTGTTGCCTACGTTACCGCG
oTM73	GTCCGTCGAGGCAACAGCTACTCGC
oTM74	CTCCTTCCAGTCGACCTGTGTGATGTC
oTM75	CACGAGTACTACAAGGAGAACTCAC
oTM76	GTCTGCGAAGATGTTCCAGCCGTCAGC
oTM77	GTGATGACGTTGACGACCTCAGC
oTM78	CTCTGGTGCGTAACCGAGTACTTCG
oTM79	TCGAGCGATGGGAAGTGATCACCAG
oTM80	ATGCTCGCGACATCGTTTCGATGTG
oTM81	GGCATGGTGGGAGAGCTGATCTCGC
oTM82	CATGCCTCCTGTCGGATCACGGACAC
oTM83	GGCACGACTGGTACGACGACGAGC
oTM86	GCTGAAGTACAGGCTCAGGTTGACAGC
oTM87	GCAACCGCGTTCTGCTCCGAGATC
oTM91	CTCGGTACCCGGGGATCCTCTAGAGCAAGCCGGTGTAACGATCTTGAGGC
oTM92	CAAGCAGAGATGGTGCCCTTGGTGACACAACCGGTCGTGACTTTTAGGG
oTM100	CGGATAGCGGGTCGACGTGCCCTTTACAGCCACCGAGAACG
oTM101	TACCTAGCCTGTCGACGGTGGCTGTCAAGTTGTTGGATAC
oTM114	CGATCATCAGTCGGAGAGTCCGACG
oTM115	GGATGGCATCGGGTCTCGACGTGGTG
oTM120	CTTGTGTGATTCTCACTCTACCGGATG
oTM121	CCCCTAGTCCAGCTCTGACCACG
oTM124	CTATCGGAAGCTAGCTGCGAGCGTAG
oTM125	CACCAACTTGGACCGGTGAACAAAGC
oTM126	CTTGCTGGTGTGATTCTCACTCTACC
oTM127	TCCGAGGTTGTGCTTGACATGCACC
oTM138	CATCACTCATGGTGTTCCTCTC
oTM139	CGCTGACTTGAGTCGTAACCAG
oTM144	TTGCTTGTGTGATTCTCACTCTAC
oTM145	AGCGACAGGACTTGTCAGACAC
oTM146	CAGCAGAATCTATCGGATCCTAG
oTM147	TGAACTCAATAGTGATTCTCCAGTC
oTM152	TGCAGCGATAGTAGACATCATC
oTM153	AGCACAGCACCCCTGTATATG
oTM154	GGATTCAACACGTTGTGACATGG
oTM155	GAGTACGAGTCACTTGACATCG
oTM156	ATAGCTAGTCGCATTATCCGTTG
oTM157	GTGATACGGTGTAACCACAAGTTC
oTM158	TTAGTGTTGTCGGACTCATCAGAG
oTM159	AATTGCGAATCGACTTGACATTAC
oTM160	CAGATGCTGTGATGAGTGTCAAC
oTM161	CAATTCGGACTTGACACTCATCAC
oTM162	CCACAGGTCTACATTCAAGTTATC
oTM163	TGAATGTCATCACCAACTTGGAC
oTM179	GATTCTCACTCTACCGGACTAGTC
oTM180	GTTTACTTGTGTGTCATCGCAGCAAG
oTM188	TTCGAGCTCGGTACCCGGGGATCCTCCATGATGTATTTGTCGCCGTGACG
oTM189	GAGATGGTGCCCTTGGTGGTCGACTCGATTCTGACTGGCACGTGGCCCCAC

oTM190	TTCGAGCTCGGTACCCGGGGATCCTCCATGACGTCCATGTGTTCTGCGGC
oTM191	GAGATGGTGCCCTTGGTGGTCGACTCGTGGTCGACATCGACGGGAGGTTG
oTM192	TTCGAGCTCGGTACCCGGGGATCCTCAGCATCACGTCTCGGATGGGCTGG
oTM193	GAGATGGTGCCCTTGGTGGTCGACTGTGTACGTGGACGACGTGGAAGACA
oTM194	TTCGAGCTCGGTACCCGGGGATCCTGTACTCGACTTTGTTTCGCGGAAAGA
oTM195	GAGATGGTGCCCTTGGTGGTCGACTCATCGACCTCGAGGAGTACATCAAG
oTM196	TTCGAGCTCGGTACCCGGGGATCCTGTGCGAGCACGTCAACCTGGTTCGGC
oTM197	GAGATGGTGCCCTTGGTGGTCGACTCAGGCTGCCGTACATCGCCGAAGAC
oTM198	TTCGAGCTCGGTACCCGGGGATCCTCCACCAGGAGCGATCTTCACGTTCT
oTM199	GAGATGGTGCCCTTGGTGGTCGACTCTCGAAGAACTGGCAGATCTCCGTG
oTM202	TCCGTAAAAACAGGTTAAAAACCG
oTM203	AAACGTTGGAATCACGCCATTCC
oTM206	CGATTGGTGAGGGTAGGAGTTTG
oTM207	TTGACGTGAATTGCAGTGATTTGC
oTM208	CGAACCACTGTGTCATCATCTC
oTM209	AGCGAGATAACTTGGACGATCAAC
oTM231	GATTCTCACTCTAGCAGTTCGTC
oTM232	GTAGAGGGCAATTTCCAATTCG
oTM237	GATTCTCACTCTACTCGACTAG
oTM238	AGGAACACACTCGACTTGACAC
oTM243	GAATGCCGAACCAAAGCTCAG
oTM244	CCGGTGAGATCAAGCGAGTAG
oTM245	AACAGTATCTATCGGAACCTAACC
oTM246	GGACTGCTAGAGTGAGAATC
oTM247	ATTTAGTGTTGTTCGGACCCATCG
oTM248	TGAAAATCCGGTGGTACAGCC
oTM251	GTCCAAGTTGGTGTGTCTCAG
oTM252	CGCAGGTAGACCCCTATTTTG
oTM253	GCAGTTCAGGCGCTATCTATC
oTM254	CAGATTCTCCGGACTTGACATTC
oTM255	ATAAGGCCCACTCGGATGAC
oTM256	CGATACCACTGGACTTGACAC
oTM257	ACCAGATCTTTAAATCTAGACTGCAGTCTCGTCGAACTGGAGTTCCTC
oTM258	ACGTCGACATCGATAAGCTTCTACTCTTCCGAAGACCCCTCG
oTM265	ACGTCGACATCGATAAGCTTCATGTTTCGGATCCTCGATGACG
oTM266	ACCAGATCTTTAAATCTAGAGGTGACCACAACAATTGCGGATCCAGCT
oTM267	ACGTCGACATCGATAAGCTTCGAATTCTGCAGCTGGATCCGCAATTG
oTM290	AGTGTTGCCGCTCATGTTTC
oTM291	GACCACCAGAGATAGCTAGG
RR-13	CTGTTGCGCTGAGGATCCCCTGTTTCGTCAC
RR-16	CCTGGTCTCGTAAGAATTCGCTACCCGGTAG
Nonspecific ^a	CTGGCTGGgTGATGGGGCGATTGTGCTCGCTGaTCGCTGGTG

^aOligonucleotide used for ParB EMSAs in Chapter 4

APPENDIX C PLASMIDS USED IN THIS STUDY

Table C-1. Plasmids used in this study.

Plasmid Name	Antibiotic Resistance	Parental Plasmid	Mycobact. Repl/Int	Insertion
pTM1	Amp ^R	pET-21a	N/A	Trixie <i>rep-His</i>
pTM8	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM9	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM10	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM11	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM12	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM14	Amp ^R , Kan ^R	pTM29	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i> ; L5 P _{left} mutant derivative
pTM29,pTM31	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	L5c ^{ts43} <i>rep</i>
pTM32	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	Bxb1 <i>rep</i>
pTM33	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	Et2Brutus <i>rep</i>
pTM34	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	Gladiator <i>rep</i>
pTM36	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	StarStuff <i>rep</i>
pTM38	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	Trixie <i>rep</i>
pTM75	Amp ^R , Kan ^R	pMH94	L5 <i>attP</i> /Int	L5 <i>rep</i>
pTM44	Gent ^R	pJV44	oriM	Δ P _{<i>hsp60</i>}
pTM48	Gent ^R	pJV44	oriM	DaVinci <i>rep</i> ; Δ P _{<i>hsp60</i>}
pTM51	Gent ^R	pJV44	oriM	DaVinci <i>rep-73</i> ; Δ P _{<i>hsp60</i>}
pTM53	Gent ^R	pJV44	oriM	phiTM46 <i>rep</i> ; Δ P _{<i>hsp60</i>}
pTM54,pTM57	Gent ^R	pJV44	oriM	phiTM46 <i>rep-73</i> ; Δ P _{<i>hsp60</i>}
pTM58	Gent ^R	pTM57	oriM	phiTM46 Δ <i>rep-73</i> ; Δ P _{<i>hsp60</i>}
pRR06	Kan ^R	pMD02	N/A	RedRock <i>parABS</i>
pJC02	Kan ^R	pET-28a	N/A	RedRock <i>parB</i>
pWN01	Kan ^R	pET-28a	N/A	Alma <i>parB</i>
pMO01	Kan ^R	pLO87	oriM	P _{<i>hsp60-mCherry</i>} ; RedRock <i>parABS</i>
pMO02	Kan ^R	pMO01	oriM	P _{<i>hsp60-mCherry</i>} ; RedRock <i>parABS</i> (<i>parA</i> *)
pMO03	Kan ^R	pMO01	oriM	P _{<i>hsp60-mCherry</i>} ; RedRock <i>parABS</i> (<i>parB</i> *)
pMO04	Kan ^R	pMO01	oriM	P _{<i>hsp60-mCherry</i>} ; RedRock <i>parABS</i> Δ <i>parS-L</i>
pMO05	Kan ^R	pMO01	oriM	P _{<i>hsp60-mCherry</i>} ; RedRock <i>parABS</i> Δ <i>parS-R</i>

BIBLIOGRAPHY

- Abeles, A.L., Friedman, S.A., and Austin, S.J. (1985). Partition of unit-copy miniplasmids to daughter cells. III. The DNA sequence and functional organization of the P1 partition region. *J Mol Biol* 185, 261-272.
- Achtman, M., and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* 6, 431-440.
- Agarwal, N., and Tyagi, A.K. (2006). Mycobacterial transcriptional signals: requirements for recognition by RNA polymerase and optimal transcriptional activity. *Nucleic Acids Res* 34, 4245-4257.
- Akimkina, T., Venien-Bryan, C., and Hodgkin, J. (2007). Isolation, characterization and complete nucleotide sequence of a novel temperate bacteriophage Min1, isolated from the nematode pathogen *Microbacterium nematophilum*. *Res Microbiol* 158, 582-590.
- Allen, L.Z., Ishoe, T., Novotny, M.A., McLean, J.S., Lasken, R.S., and Williamson, S.J. (2011). Single virus genomics: a new tool for virus discovery. *PLoS One* 6, e17722.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Andersson, A.F., and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320, 1047-1050.
- Arbolea, S., Watkins, C., Stanton, C., and Ross, R.P. (2016). Gut Bifidobacteria Populations in Human Health and Aging. *Front Microbiol* 7, 1204.
- Asadulghani, M., Ogura, Y., Ooka, T., Itoh, T., Sawaguchi, A., Iguchi, A., Nakayama, K., and Hayashi, T. (2009). The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog* 5, e1000408.
- Austin, S., and Abeles, A. (1983). Partition of unit-copy miniplasmids to daughter cells. II. The partition region of miniplasmid P1 encodes an essential protein and a centromere-like site at which it acts. *J Mol Biol* 169, 373-387.
- Austin, S., and Nordstrom, K. (1990). Partition-mediated incompatibility of bacterial plasmids. *Cell* 60, 351-354.

- Bachrach, G., Colston, M.J., Bercovier, H., Bar-Nir, D., Anderson, C., and Papavinasasundaram, K.G. (2000). A new single-copy mycobacterial plasmid, pMF1, from *Mycobacterium fortuitum* which is compatible with the pAL5000 replicon. *Microbiology* 146 (Pt 2), 297-303.
- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37, W202-208.
- Bailone, A., and Devoret, R. (1978). Isolation of ultravirulent mutants of phage lambda. *Virology* 84, 547-550.
- Baltz, R.H. (2012). *Streptomyces* temperate bacteriophage integration systems for stable genetic engineering of actinomycetes (and other organisms). *J Ind Microbiol Biotechnol* 39, 661-672.
- Bandhu, A., Ganguly, T., Jana, B., Mondal, R., and Sau, S. (2010). Regions and residues of an asymmetric operator DNA interacting with the monomeric repressor of temperate mycobacteriophage L1. *Biochemistry* 49, 4235-4243.
- Barka, E.A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Meier-Kolthoff, J.P., Klenk, H.P., Clement, C., Ouhdouch, Y., and van Wezel, G.P. (2016). Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev* 80, 1-43.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709-1712.
- Bashyam, M.D., Kaushal, D., Dasgupta, S.K., and Tyagi, A.K. (1996). A study of mycobacterial transcriptional apparatus: identification of novel features in promoter elements. *J Bacteriol* 178, 4847-4853.
- Bebeacua, C., Bron, P., Lai, L., Vegge, C.S., Brondsted, L., Spinelli, S., Campanacci, V., Veessler, D., van Heel, M., and Cambillau, C. (2010). Structure and molecular assignment of lactococcal phage TP901-1 baseplate. *J Biol Chem* 285, 39079-39086.
- Belitsky, B.R., and Sonenshein, A.L. (2011). Roadblock repression of transcription by *Bacillus subtilis* CodY. *J Mol Biol* 411, 729-743.
- Bell, C.E., Frescura, P., Hochschild, A., and Lewis, M. (2000). Crystal structure of the lambda repressor C-terminal domain provides a model for cooperative operator binding. *Cell* 101, 801-811.

- Bell, C.E., and Lewis, M. (2001). Crystal structure of the lambda repressor C-terminal domain octamer. *J Mol Biol* 314, 1127-1136.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 573-580.
- Bergh, O., Borsheim, K.Y., Bratbak, G., and Haldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature* 340, 467-468.
- Berngruber, T.W., Weissing, F.J., and Gandon, S. (2010). Inhibition of superinfection and the evolution of viral latency. *J Virol* 84, 10200-10208.
- Bertani, L.E. (1971). Stabilization of P2 tandem double lysogens by int mutations in the prophage. *Virology* 46, 426-436.
- Bertrand, C., Thibessard, A., Bruand, C., Lecointe, F., and Leblond, P. (2019). Bacterial NHEJ: A never ending story. *Mol Microbiol*.
- Bibb, L.A., and Hatfull, G.F. (2002). Integration and excision of the Mycobacterium tuberculosis prophage-like element, phiRv1. *Mol Microbiol* 45, 1515-1526.
- Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci U S A* 104, 3312-3317.
- Bobay, L.M., Rocha, E.P., and Touchon, M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* 30, 737-751.
- Bobay, L.M., Touchon, M., and Rocha, E.P. (2014). Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci U S A* 111, 12127-12132.
- Bondy-Denomy, J., Qian, J., Westra, E.R., Buckling, A., Guttman, D.S., Davidson, A.R., and Maxwell, K.L. (2016). Prophages mediate defense against phage infection through diverse mechanisms. *ISME J* 10, 2854-2866.
- Bonnet, J., Subsoontorn, P., and Endy, D. (2012). Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc Natl Acad Sci U S A* 109, 8884-8889.
- Borodovsky, M., Mills, R., Besemer, J., and Lomsadze, A. (2003). Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics Chapter 4*, Unit4 5.
- Bossi, L., Fuentes, J.A., Mora, G., and Figueroa-Bossi, N. (2003). Prophage contribution to bacterial population dynamics. *J Bacteriol* 185, 6467-6471.

- Bottacini, F., Morrissey, R., Roberts, R.J., James, K., van Breen, J., Egan, M., Lambert, J., van Limpt, K., Knol, J., Motherway, M.O., *et al.* (2017). Comparative genome and methylome analysis reveals restriction/modification system diversity in the gut commensal *Bifidobacterium breve*. *Nucleic Acids Res.*
- Bottacini, F., O'Connell Motherway, M., Kuczynski, J., O'Connell, K.J., Serafini, F., Duranti, S., Milani, C., Turrone, F., Lugli, G.A., Zomer, A., *et al.* (2014). Comparative genomics of the *Bifidobacterium breve* taxon. *BMC Genomics* 15, 170.
- Bourhy, P., Frangeul, L., Couve, E., Glaser, P., Saint Girons, I., and Picardeau, M. (2005). Complete nucleotide sequence of the LE1 prophage from the spirochete *Leptospira biflexa* and characterization of its replication and partition functions. *J Bacteriol* 187, 3931-3940.
- Bourn, W.R., Jansen, Y., Stutz, H., Warren, R.M., Williamson, A.L., and van Helden, P.D. (2007). Creation and characterisation of a high-copy-number version of the pAL5000 mycobacterial replicon. *Tuberculosis (Edinb)* 87, 481-488.
- Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 99, 14250-14255.
- Briner, A.E., Lugli, G.A., Milani, C., Duranti, S., Turrone, F., Gueimonde, M., Margolles, A., van Sinderen, D., Ventura, M., and Barrangou, R. (2015). Occurrence and Diversity of CRISPR-Cas Systems in the Genus *Bifidobacterium*. *PLoS One* 10, e0133661.
- Bronson, M.J., and Levine, M. (1971). Virulent mutants of bacteriophage p22.I. Isolation and genetic analysis. *J Virol* 7, 559-568.
- Broussard, G.W., Oldfield, L.M., Villanueva, V.M., Lunt, B.L., Shine, E.E., and Hatfull, G.F. (2013). Integration-dependent bacteriophage immunity provides insights into the evolution of genetic switches. *Mol Cell* 49, 237-248.
- Brown, K.L., Sarkis, G.J., Wadsworth, C., and Hatfull, G.F. (1997). Transcriptional silencing by the mycobacteriophage L5 repressor. *EMBO J* 16, 5914-5921.
- Browning, D.F., and Busby, S.J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* 2, 57-65.
- Brussaard, C.P., Marie, D., and Bratbak, G. (2000). Flow cytometric detection of viruses. *J Virol Methods* 85, 175-182.
- Brussow, H. (2001). Phages of dairy bacteria. *Annu Rev Microbiol* 55, 283-303.

- Brussow, H., Canchaya, C., and Hardt, W.D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* 68, 560-602, table of contents.
- Brussow, H., and Hendrix, R.W. (2002). Phage genomics: small is beautiful. *Cell* 108, 13-16.
- Bruttin, A., Desiere, F., d'Amico, N., Guerin, J.P., Sidoti, J., Huni, B., Lucchini, S., and Brussow, H. (1997a). Molecular ecology of *Streptococcus thermophilus* bacteriophage infections in a cheese factory. *Appl Environ Microbiol* 63, 3144-3150.
- Bruttin, A., Desiere, F., Lucchini, S., Foley, S., and Brussow, H. (1997b). Characterization of the lysogeny DNA module from the temperate *Streptococcus thermophilus* bacteriophage phi Sfi21. *Virology* 233, 136-148.
- Bryan, A., Youngster, I., and McAdam, A.J. (2015). Shiga Toxin Producing *Escherichia coli*. *Clin Lab Med* 35, 247-272.
- Campbell, A. (1994). Comparative molecular biology of lambdoid phages. *Annu Rev Microbiol* 48, 193-222.
- Carlson, P.A., and Koudelka, G.B. (1994). Expression, purification, and functional characterization of the carboxyl-terminal domain fragment of bacteriophage 434 repressor. *J Bacteriol* 176, 6907-6914.
- Carmelo, E., Barilla, D., Golovanov, A.P., Lian, L.Y., Derome, A., and Hayes, F. (2005). The unstructured N-terminal tail of ParG modulates assembly of a quaternary nucleoprotein complex in transcription repression. *J Biol Chem* 280, 28683-28691.
- Casey, E., Mahony, J., O'Connell-Motherway, M., Bottacini, F., Cornelissen, A., Neve, H., Heller, K.J., Noben, J.P., Dal Bello, F., and van Sinderen, D. (2014). Molecular characterization of three *Lactobacillus delbrueckii* subsp. *bulgaricus* phages. *Appl Environ Microbiol* 80, 5623-5635.
- Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* 49, 277-300.
- Casjens, S., Hatfull, G.F., and Hendrix, R.W. (1992). Evolution of dsDNA tailed-bacteriophage genomes. *Seminars in Virology* 3, 383-397.
- Casjens, S.R., Gilcrease, E.B., Huang, W.M., Bunny, K.L., Pedulla, M.L., Ford, M.E., Houtz, J.M., Hatfull, G.F., and Hendrix, R.W. (2004). The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J Bacteriol* 186, 1818-1832.

- Casjens, S.R., and Hendrix, R.W. (2015). Bacteriophage lambda: Early pioneer and still relevant. *Virology* 479-480, 310-330.
- Chen, F., and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl Environ Microbiol* 68, 2589-2594.
- Cheng, L., Marinelli, L.J., Grosset, N., Fitz-Gibbon, S.T., Bowman, C.A., Dang, B.Q., Russell, D.A., Jacobs-Sera, D., Shi, B., Pellegrini, M., *et al.* (2018). Complete genomic sequences of *Propionibacterium freudenreichii* phages from Swiss cheese reveal greater diversity than *Cutibacterium* (formerly *Propionibacterium*) *acnes* phages. *BMC Microbiol* 18, 19.
- Chew, D.S., Leung, M.Y., and Choi, K.P. (2007). AT excursion: a new approach to predict replication origins in viral genomes by locating AT-rich regions. *BMC Bioinformatics* 8, 163.
- Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.L., and Brussow, H. (2004). Phage-host interaction: an ecological perspective. *J Bacteriol* 186, 3677-3686.
- Chithambaram, S., Prabhakaran, R., and Xia, X. (2014a). Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. *Mol Biol Evol* 31, 1606-1617.
- Chithambaram, S., Prabhakaran, R., and Xia, X. (2014b). The effect of mutation and selection on codon adaptation in *Escherichia coli* bacteriophage. *Genetics* 197, 301-315.
- Chopin, A., Bolotin, A., Sorokin, A., Ehrlich, S.D., and Chopin, M. (2001). Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. *Nucleic Acids Res* 29, 644-651.
- Ciampi, M.S. (2006). Rho-dependent terminators and transcription termination. *Microbiology* 152, 2515-2528.
- Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422-1423.
- Colin, J., Candelli, T., Porrua, O., Boulay, J., Zhu, C., Lacroute, F., Steinmetz, L.M., and Libri, D. (2014). Roadblock termination by *reb1p* restricts cryptic and readthrough transcription. *Mol Cell* 56, 667-680.
- Coyne, M.J., Weinacht, K.G., Krinos, C.M., and Comstock, L.E. (2003). *Mpi* recombinase globally modulates the surface architecture of a human commensal bacterium. *Proc Natl Acad Sci U S A* 100, 10446-10451.

- Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R.W., and Hatfull, G.F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics* 12, 395.
- Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res* 14, 1188-1190.
- Csuros, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910-1912.
- D’Herelle, F. (1917). Sur un microbe invisible antagoniste des bacillus dysentérique. *Acad Sci Paris* 165, 373–375.
- Da Silva, M., and Upton, C. (2005). Using purine skews to predict genes in AT-rich poxviruses. *BMC Genomics* 6, 22.
- Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5, e11147.
- De Anda, J., Poteete, A.R., and Sauer, R.T. (1983). P22 c2 repressor. Domain structure and function. *J Biol Chem* 258, 10536-10542.
- Dedrick, R.M., Jacobs-Sera, D., Bustamante, C.A., Garlena, R.A., Mavrich, T.N., Pope, W.H., Reyes, J.C., Russell, D.A., Adair, T., Alvey, R., *et al.* (2017a). Prophage-mediated defence against viral attack and viral counter-defence. *Nat Microbiol* 2, 16251.
- Dedrick, R.M., Mavrich, T.N., Ng, W.L., Cervantes Reyes, J.C., Olm, M.R., Rush, R.E., Jacobs-Sera, D., Russell, D.A., and Hatfull, G.F. (2016). Function, expression, specificity, diversity and incompatibility of actinobacteriophage parABS systems. *Mol Microbiol* 101, 625-644.
- Dedrick, R.M., Mavrich, T.N., Ng, W.L., and Hatfull, G.F. (2017b). Expression and evolutionary patterns of mycobacteriophage D29 and its temperate close relatives. *BMC Microbiol* 17, 225.
- Degnan, P.H., Michalowski, C.B., Babic, A.C., Cordes, M.H., and Little, J.W. (2007). Conservation and diversity in the immunity regions of wild phages with the immunity specificity of phage lambda. *Mol Microbiol* 64, 232-244.
- del Solar, G., Giraldo, R., Ruiz-Echevarria, M.J., Espinosa, M., and Diaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* 62, 434-464.
- Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27, 4636-4641.

- Denes, T., Vongkamjan, K., Ackermann, H.W., Moreno Switt, A.I., Wiedmann, M., and den Bakker, H.C. (2014). Comparative genomic and morphological analyses of *Listeria* phages isolated from farm environments. *Appl Environ Microbiol* 80, 4616-4625.
- Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P., and Sullivan, M.B. (2014). Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* 513, 242-245.
- Dmowski, M., Sitkiewicz, I., and Ceglowski, P. (2006). Characterization of a novel partition system encoded by the delta and omega genes from the streptococcal plasmid pSM19035. *J Bacteriol* 188, 4362-4372.
- Dobbins, A.T., George, M., Jr., Basham, D.A., Ford, M.E., Houtz, J.M., Pedulla, M.L., Lawrence, J.G., Hatfull, G.F., and Hendrix, R.W. (2004). Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J Bacteriol* 186, 1933-1944.
- Donnelly-Wu, M.K., Jacobs, W.R., Jr., and Hatfull, G.F. (1993). Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria. *Mol Microbiol* 7, 407-417.
- Donner, A.L., Paa, K., and Koudelka, G.B. (1998). Carboxyl-terminal domain dimer interface mutant 434 repressors have altered dimerization and DNA binding specificities. *J Mol Biol* 283, 931-946.
- Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359.
- Durmaz, E., Madsen, S.M., Israelsen, H., and Klaenhammer, T.R. (2002). *Lactococcus lactis* lytic bacteriophages of the P335 group are inhibited by overexpression of a truncated CI repressor. *J Bacteriol* 184, 6532-6544.
- Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., *et al.* (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 5, 4498.
- Ebersbach, G., and Gerdes, K. (2005). Plasmid segregation mechanisms. *Annu Rev Genet* 39, 453-479.
- Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol* 3, 504-510.

- Fan, X., Xie, L., Li, W., and Xie, J. (2014). Prophage-like elements present in *Mycobacterium* genomes. *BMC Genomics* 15, 243.
- Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., and Herskovits, A.A. (2015). A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol* 13, 641-650.
- Fernandez, L., Gonzalez, S., Quiles-Puchalt, N., Gutierrez, D., Penades, J.R., Garcia, P., and Rodriguez, A. (2018). Lysogenization of *Staphylococcus aureus* RN450 by phages varphi11 and varphi80alpha leads to the activation of the SigB regulon. *Sci Rep* 8, 12662.
- Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., *et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* 260, 500-507.
- Fitch, W.M., and Ayala, F.J. (1994). Tempo and mode in evolution. *Proc Natl Acad Sci U S A* 91, 6717-6720.
- Ford, M.E., Sarkis, G.J., Belanger, A.E., Hendrix, R.W., and Hatfull, G.F. (1998). Genome structure of mycobacteriophage D29: implications for phage evolution. *J Mol Biol* 279, 143-164.
- Fothergill, T.J., Barilla, D., and Hayes, F. (2005). Protein diversity confers specificity in plasmid segregation. *J Bacteriol* 187, 2651-2661.
- Frenzel, A., Schirrmann, T., and Hust, M. (2016). Phage display-derived human antibodies in clinical development and therapy. *MAbs* 8, 1177-1194.
- Fukuda, S., Toh, H., Hase, K., Oshima, K., Nakanishi, Y., Yoshimura, K., Tobe, T., Clarke, J.M., Topping, D.L., Suzuki, T., *et al.* (2011). Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* 469, 543-547.
- Ganguly, T., Bandhu, A., Chattoraj, P., Chanda, P.K., Das, M., Mandal, N.C., and Sau, S. (2007). Repressor of temperate mycobacteriophage L1 harbors a stable C-terminal domain and binds to different asymmetric operator DNAs with variable affinity. *Virol J* 4, 64.
- Gavigan, J.A., Ainsa, J.A., Perez, E., Otal, I., and Martin, C. (1997). Isolation by genetic labeling of a new mycobacterial plasmid, pJAZ38, from *Mycobacterium fortuitum*. *J Bacteriol* 179, 4115-4122.

- Gerdes, K., Moller-Jensen, J., and Bugge Jensen, R. (2000). Plasmid and chromosome partitioning: surprises from phylogeny. *Mol Microbiol* 37, 455-466.
- Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6, 343-345.
- Gomez-Escribano, J.P., Castro, J.F., Razmilic, V., Chandra, G., Andrews, B., Asenjo, J.A., and Bibb, M.J. (2015). The *Streptomyces leeuwenhoekii* genome: de novo sequencing and assembly in single contigs of the chromosome, circular plasmid pSLE1 and linear plasmid pSLE2. *BMC Genomics* 16, 485.
- Gordon, B.R., Imperial, R., Wang, L., Navarre, W.W., and Liu, J. (2008). Lsr2 of *Mycobacterium* represents a novel class of H-NS-like proteins. *J Bacteriol* 190, 7052-7059.
- Gottesman, M.E., and Weisberg, R.A. (2004). Little lambda, who made thee? *Microbiol Mol Biol Rev* 68, 796-813.
- Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27, 221-224.
- Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* 45, D491-D498.
- Greg Finak, M.J. (2011). flowWorkspace: Infrastructure for representing and interacting with the gated cytometry. R package version 3.26.2.
- Gregory, A.C., Solonenko, S.A., Ignacio-Espinoza, J.C., LaButti, K., Copeland, A., Sudek, S., Maitland, A., Chittick, L., Dos Santos, F., Weitz, J.S., *et al.* (2016). Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 17, 930.
- Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 26, 2286-2290.
- Grigoriev, A. (1999). Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus research* 60, 1-19.
- Grindley, N.D., Whiteson, K.L., and Rice, P.A. (2006). Mechanisms of site-specific recombination. *Annu Rev Biochem* 75, 567-605.

- Grose, J.H., and Casjens, S.R. (2014). Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* 468-470, 421-443.
- Haaber, J., Leisner, J.J., Cohn, M.T., Catalan-Moreno, A., Nielsen, J.B., Westh, H., Penades, J.R., and Ingmer, H. (2016). Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat Commun* 7, 13333.
- Hahne, F., LeMeur, N., Brinkman, R.R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., and Gentleman, R. (2009). flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* 10, 106.
- Hammerl, J.A., Jackel, C., Lanka, E., Roschanski, N., and Hertwig, S. (2016). Binding Specificities of the Telomere Phage varphiKO2 Prophage Repressor CB and Lytic Repressor Cro. *Viruses* 8.
- Hammerl, J.A., Klevanskaa, K., Strauch, E., and Hertwig, S. (2014). Complete Nucleotide Sequence of pVv01, a P1-Like Plasmid Prophage of *Vibrio vulnificus*. *Genome Announc* 2.
- Hammerl, J.A., Roschanski, N., Lurz, R., Johne, R., Lanka, E., and Hertwig, S. (2015). The Molecular Switch of Telomere Phages: High Binding Specificity of the PY54 Cro Lytic Repressor to a Single Operator Site. *Viruses* 7, 2771-2793.
- Hanauer, D.I., Graham, M.J., Sea, P., Betancur, L., Bobrownicki, A., Cresawn, S.G., Garlena, R.A., Jacobs-Sera, D., Kaufmann, N., Pope, W.H., *et al.* (2017). An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. *Proc Natl Acad Sci U S A* 114, 13531-13536.
- Hanauer, D.I., Jacobs-Sera, D., Pedulla, M.L., Cresawn, S.G., Hendrix, R.W., and Hatfull, G.F. (2006). Inquiry learning. Teaching scientific inquiry. *Science* 314, 1880-1881.
- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99-104.
- Hatfull, G.F. (2010). Mycobacteriophages: genes and genomes. *Annu Rev Microbiol* 64, 331-356.
- Hatfull, G.F. (2012). The secret lives of mycobacteriophages. *Adv Virus Res* 82, 179-288.
- Hatfull, G.F. (2014). Molecular Genetics of Mycobacteriophages. *Microbiol Spectr* 2, 1-36.
- Hatfull, G.F. (2018). Mycobacteriophages. *Microbiol Spectr* 6.

- Hatfull, G.F., and Hendrix, R.W. (2011). Bacteriophages and their genomes. *Curr Opin Virol* 1, 298-303.
- Hatfull, G.F., Jacobs-Sera, D., Lawrence, J.G., Pope, W.H., Russell, D.A., Ko, C.C., Weber, R.J., Patel, M.C., Germane, K.L., Edgar, R.H., *et al.* (2010). Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol* 397, 119-143.
- Hatfull, G.F., and Sarkis, G.J. (1993). DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Mol Microbiol* 7, 395-405.
- Haugen, S.P., Ross, W., and Gourse, R.L. (2008). Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat Rev Microbiol* 6, 507-519.
- Hauser, M., Mayer, C.E., and Soding, J. (2013). kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* 14, 248.
- Hayes, F., and Austin, S.J. (1993). Specificity determinants of the P1 and P7 plasmid centromere analogs. *Proc Natl Acad Sci U S A* 90, 9228-9232.
- Heinrich, J., Velleman, M., and Schuster, H. (1995). The tripartite immunity system of phages P1 and P7. *FEMS Microbiol Rev* 17, 121-126.
- Heinzel, T., Velleman, M., and Schuster, H. (1992). C1 repressor of phage P1 is inactivated by noncovalent binding of P1 Coi protein. *J Biol Chem* 267, 4183-4188.
- Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* 96, 2192-2197.
- Hershey, A.D. (1971). *The Bacteriophage lambda* (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory).
- Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* 36, 39-56.
- Hertwig, S., Klein, I., Schmidt, V., Beck, S., Hammerl, J.A., and Appel, B. (2003). Sequence analysis of the genome of the temperate *Yersinia enterocolitica* phage PY54. *J Mol Biol* 331, 605-622.

- Highton, P.J., Chang, Y., and Myers, R.J. (1990). Evidence for the exchange of segments between genomes during the evolution of lambdoid bacteriophages. *Mol Microbiol* 4, 1329-1340.
- Hofer, B., Ruge, M., and Dreiseikelmann, B. (1995). The superinfection exclusion gene (sieA) of bacteriophage P22: identification and overexpression of the gene and localization of the gene product. *J Bacteriol* 177, 3080-3086.
- Howe, M.M., and Bade, E.G. (1975). Molecular biology of bacteriophage mu. *Science* 190, 624-632.
- Huang, J., Ghosh, P., Hatfull, G.F., and Hong, Y. (2011). Successive and targeted DNA integrations in the *Drosophila* genome by Bxb1 and phiC31 integrases. *Genetics* 189, 391-395.
- Huerta-Cepas, J., Dopazo, J., and Gabaldon, T. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11, 24.
- Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., HERNSDORF, A.W., Amano, Y., Ise, K., *et al.* (2016). A new view of the tree of life. *Nat Microbiol* 1, 16048.
- Iandolo, J.J., Worrell, V., Groicher, K.H., Qian, Y., Tian, R., Kenton, S., Dorman, A., Ji, H., Lin, S., Loh, P., *et al.* (2002). Comparative analysis of the genomes of the temperate bacteriophages phi 11, phi 12 and phi 13 of *Staphylococcus aureus* 8325. *Gene* 289, 109-118.
- Ignacio-Espinoza, J.C., Solonenko, S.A., and Sullivan, M.B. (2013). The global virome: not as big as we thought? *Curr Opin Virol* 3, 566-571.
- Ingham, C.J., Crombie, H.J., Bruton, C.J., Chater, K.F., Hartley, N.M., Murphy, G.J., and Smith, M.C. (1993). Multiple novel promoters from the early region in the *Streptomyces* temperate phage phi C31 are activated during lytic development. *Mol Microbiol* 9, 1267-1274.
- Ingham, C.J., Owen, C.E., Wilson, S.E., Hunter, I.S., and Smith, M.C. (1994). An operator associated with autoregulation of the repressor gene in actinophage phiC31 is found in highly conserved copies in intergenic regions in the phage genome. *Nucleic Acids Res* 22, 821-827.
- Ingham, C.J., and Smith, M.C. (1992). Transcription map of the early region of the *Streptomyces* bacteriophage phi C31. *Gene* 122, 77-84.

- Iro, M., Klein, R., Galos, B., Baranyi, U., Rossler, N., and Witte, A. (2007). The lysogenic region of virus phiCh1: identification of a repressor-operator system and determination of its activity in halophilic Archaea. *Extremophiles* 11, 383-396.
- Jacobs-Sera, D., Marinelli, L.J., Bowman, C., Broussard, G.W., Guerrero Bustamante, C., Boyle, M.M., Petrova, Z.O., Dedrick, R.M., Pope, W.H., Science Education Alliance Phage Hunters Advancing, G., *et al.* (2012). On the nature of mycobacteriophage diversity and host preference. *Virology* 434, 187-201.
- Jain, S., and Hatfull, G.F. (2000). Transcriptional regulation and immunity in mycobacteriophage Bxb1. *Mol Microbiol* 38, 971-985.
- Jamet, A., Touchon, M., Ribeiro-Goncalves, B., Carrico, J.A., Charbit, A., Nassif, X., Ramirez, M., and Rocha, E.P.C. (2017). A widespread family of polymorphic toxins encoded by temperate phages. *BMC Biol* 15, 75.
- Jensen, R.B., and Gerdes, K. (1995). Programmed cell death in bacteria: proteic plasmid stabilization systems. *Mol Microbiol* 17, 205-210.
- Johnson, R.C. (2015). Site-specific DNA Inversion by Serine Recombinases. *Microbiol Spectr* 3, MDNA3-0047-2014.
- Jordan, T.C., Burnett, S.H., Carson, S., Caruso, S.M., Clase, K., DeJong, R.J., Dennehy, J.J., Denver, D.R., Dunbar, D., Elgin, S.C., *et al.* (2014). A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *MBio* 5, e01051-01013.
- Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., and Hendrix, R.W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 299, 27-51.
- Kameyama, L., Fernandez, L., Calderon, J., Ortiz-Rojas, A., and Patterson, T.A. (1999). Characterization of wild lambdoid bacteriophages: detection of a wide distribution of phage immunity groups and identification of a nus-dependent, nonlambdoid phage group. *Virology* 263, 100-111.
- Karlsson, J.L., Cardoso-Palacios, C., Nilsson, A.S., and Haggard-Ljungquist, E. (2006). Evolution of immunity and host chromosome integration site of P2-like coliphages. *J Bacteriol* 188, 3923-3935.

- Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059-3066.
- Keen, E.C. (2015). A century of phage research: bacteriophages and the shaping of modern biology. *Bioessays* 37, 6-9.
- Kim, A.I., Ghosh, P., Aaron, M.A., Bibb, L.A., Jain, S., and Hatfull, G.F. (2003). Mycobacteriophage Bxb1 integrates into the Mycobacterium smegmatis groEL1 gene. *Mol Microbiol* 50, 463-473.
- Kim, J., Ochoa, M.T., Krutzik, S.R., Takeuchi, O., Uematsu, S., Legaspi, A.J., Brightbill, H.D., Holland, D., Cunliffe, W.J., Akira, S., *et al.* (2002). Activation of toll-like receptor 2 in acne triggers inflammatory cytokine responses. *J Immunol* 169, 1535-1541.
- Kim, M.S., and Bae, J.W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J* 12, 1127-1141.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624-626.
- Klumpp, J., Fouts, D.E., and Sozhamannan, S. (2012). Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage* 2, 190-199.
- Klumpp, J., and Loessner, M.J. (2013). Listeria phages: Genomes, evolution, and application. *Bacteriophage* 3, e26861.
- Klyczek, K.K., Bonilla, J.A., Jacobs-Sera, D., Adair, T.L., Afram, P., Allen, K.G., Archambault, M.L., Aziz, R.M., Bagnasco, F.G., Ball, S.L., *et al.* (2017). Tales of diversity: Genomic and morphological characteristics of forty-six *Arthrobacter* phages. *PLoS One* 12, e0180517.
- Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobian-Guemes, A.G., Coutinho, F.H., Dinsdale, E.A., Felts, B., Furby, K.A., *et al.* (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466-470.
- Ko, C.C., and Hatfull, G.F. (2018). Mycobacteriophage Fruitloop gp52 inactivates Wag31 (DivIVA) to prevent heterotypic superinfection. *Mol Microbiol* 108, 443-460.
- Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361, 1929-1940.
- Konstantinidis, K.T., and Tiedje, J.M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102, 2567-2572.

- Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* 479-480, 2-25.
- Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* 36, 6688-6719.
- Kreuzer, K.N. (2013). DNA damage responses in prokaryotes: regulating gene expression, modulating growth patterns, and manipulating replication forks. *Cold Spring Harb Perspect Biol* 5, a012674.
- Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23, 1026-1028.
- Krupovic, M., Prangishvili, D., Hendrix, R.W., and Bamford, D.H. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol Mol Biol Rev* 75, 610-635.
- Kuhn, J.H., Wolf, Y.I., Krupovic, M., Zhang, Y., Maes, P., Dolja, V.V., and Koonin, E.V. (2019). Classify viruses — the gain is worth the pain. *Nature* 566, 318-320.
- Kwan, T., Liu, J., DuBow, M., Gros, P., and Pelletier, J. (2005). The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc Natl Acad Sci U S A* 102, 5174-5179.
- Kwan, T., Liu, J., Dubow, M., Gros, P., and Pelletier, J. (2006). Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. *J Bacteriol* 188, 1184-1187.
- Labidi, A., David, H.L., and Roulland-Dussoix, D. (1985). Restriction endonuclease mapping and cloning of *Mycobacterium fortuitum* var. *fortuitum* plasmid pAL5000. *Ann Inst Pasteur Microbiol* (1985) 136B, 209-215.
- Labidi, A., Mardis, E., Roe, B.A., and Wallace, R.J., Jr. (1992). Cloning and DNA sequence of the *Mycobacterium fortuitum* var *fortuitum* plasmid pAL5000. *Plasmid* 27, 130-140.
- Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 8, 317-327.
- Landy, A. (1989). Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem* 58, 913-949.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359.

- Larsen, R.A., Cusumano, C., Fujioka, A., Lim-Fong, G., Patterson, P., and Pogliano, J. (2007). Treadmilling of a prokaryotic tubulin-like protein, TubZ, required for plasmid stability in *Bacillus thuringiensis*. *Genes Dev* 21, 1340-1352.
- Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 32, 11-16.
- Lawrence, J.G., Hatfull, G.F., and Hendrix, R.W. (2002). Imbrolios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J Bacteriol* 184, 4891-4905.
- Lawrence, J.G., Hendrix, R.W., and Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol* 9, 535-540.
- Lederberg, E.M. (1951). Lysogenicity in *E. coli* K-12. *Genetics* 36, 560.
- Lederberg, E.M., and Lederberg, J. (1953). Genetic Studies of Lysogenicity in *Escherichia Coli*. *Genetics* 38, 51-64.
- Lee, D.J., Minchin, S.D., and Busby, S.J. (2012). Activating transcription in bacteria. *Annu Rev Microbiol* 66, 125-152.
- Lee, M.H., Pascopella, L., Jacobs, W.R., Jr., and Hatfull, G.F. (1991). Site-specific integration of mycobacteriophage L5: integration-proficient vectors for *Mycobacterium smegmatis*, *Mycobacterium tuberculosis*, and bacille Calmette-Guerin. *Proc Natl Acad Sci U S A* 88, 3111-3115.
- Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res* 38, D57-61.
- Lewis, M. (2011). A tale of two repressors. *J Mol Biol* 409, 14-27.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* 25, 762-777.
- Lippi, D., Gotuzzo, E., and Caini, S. (2016). Cholera. *Microbiol Spectr* 4.
- Little, J.W. (2010). Evolution of complex gene regulatory circuits by addition of refinements. *Curr Biol* 20, R724-734.

- Livny, J., Yamaichi, Y., and Waldor, M.K. (2007). Distribution of centromere-like parS sites in bacteria: insights from comparative genomics. *J Bacteriol* *189*, 8693-8703.
- Lobocka, M.B., Rose, D.J., Plunkett, G., 3rd, Rusin, M., Samojedny, A., Lehnherr, H., Yarmolinsky, M.B., and Blattner, F.R. (2004). Genome of bacteriophage P1. *J Bacteriol* *186*, 7032-7068.
- Lobry, J.R. (1996). A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie* *78*, 323-326.
- Loenen, W.A., Dryden, D.T., Raleigh, E.A., Wilson, G.G., and Murray, N.E. (2014). Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res* *42*, 3-19.
- Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* *25*, 955-964.
- Loytynoja, A., and Goldman, N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* *11*, 579.
- Lucks, J.B., Nelson, D.R., Kudla, G.R., and Plotkin, J.B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS computational biology* *4*, e1000001.
- Lugli, G.A., Milani, C., Mancabelli, L., van Sinderen, D., and Ventura, M. (2016a). MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol Lett* *363*.
- Lugli, G.A., Milani, C., Turrone, F., Tremblay, D., Ferrario, C., Mancabelli, L., Duranti, S., Ward, D.V., Ossiprandi, M.C., Moineau, S., *et al.* (2016b). Prophages of the genus *Bifidobacterium* as modulating agents of the infant gut microbiota. *Environ Microbiol* *18*, 2196-2213.
- Lunt, B.L., and Hatfull, G.F. (2016). Brujita Integrase: A Simple, Arm-Less, Directionless, and Promiscuous Tyrosine Integrase System. *J Mol Biol* *428*, 2289-2306.
- MacDonald, I.C., and Deans, T.L. (2016). Tools and applications in synthetic biology. *Adv Drug Deliv Rev* *105*, 20-34.
- Madsen, P.L., and Hammer, K. (1998). Temporal transcription of the lactococcal temperate phage TP901-1 and DNA sequence of the early promoter region. *Microbiology* *144* (Pt 8), 2203-2215.
- Maniloff, J., and Ackermann, H.W. (1998). Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. *Arch Virol* *143*, 2051-2063.

- Mann, N.H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev* 27, 17-34.
- Manrique, P., Bolduc, B., Walk, S.T., van der Oost, J., de Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proc Natl Acad Sci U S A* 113, 10400-10405.
- Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., *et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39, D225-229.
- Mardanov, A.V., and Ravin, N.V. (2007). The antirepressor needed for induction of linear plasmid-prophage N15 belongs to the SOS regulon. *J Bacteriol* 189, 6333-6338.
- Marinelli, L.J., Fitz-Gibbon, S., Hayes, C., Bowman, C., Inkeles, M., Loncaric, A., Russell, D.A., Jacobs-Sera, D., Cokus, S., Pellegrini, M., *et al.* (2012). *Propionibacterium acnes* bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio* 3.
- Marinelli, L.J., Piuri, M., Swigonova, Z., Balachandran, A., Oldfield, L.M., van Kessel, J.C., and Hatfull, G.F. (2008). BRED: a simple and powerful tool for constructing mutant and recombinant bacteriophage genomes. *PLoS One* 3, e3957.
- Martinson, J.T., Radman, M., and Petit, M.A. (2008). The lambda red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. *PLoS Genet* 4, e1000065.
- Massad, T., Skaar, K., Nilsson, H., Damberg, P., Henriksson-Peltola, P., Haggard-Ljungquist, E., Hogbom, M., and Stenmark, P. (2010). Crystal structure of the P2 C-repressor: a binder of non-palindromic direct DNA repeats. *Nucleic Acids Res* 38, 7778-7790.
- Matsushiro, A., Sato, K., Miyamoto, H., Yamamura, T., and Honda, T. (1999). Induction of prophages of enterohemorrhagic *Escherichia coli* O157:H7 with norfloxacin. *J Bacteriol* 181, 2257-2260.
- Mavrich, T.N., Casey, E., Oliveira, J., Bottacini, F., James, K., Franz, C., Lugli, G.A., Neve, H., Ventura, M., Hatfull, G.F., *et al.* (2018). Characterization and induction of prophages in human gut-associated *Bifidobacterium* hosts. *Sci Rep* 8, 12772.
- Mavrich, T.N., and Hatfull, G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol* 2, 17112.

- McLean, M.J., Wolfe, K.H., and Devine, K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* 47, 691-696.
- McNair, K., Bailey, B.A., and Edwards, R.A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 28, 614-618.
- Mediavilla, J., Jain, S., Kriakov, J., Ford, M.E., Duda, R.L., Jacobs, W.R., Jr., Hendrix, R.W., and Hatfull, G.F. (2000). Genome organization and characterization of mycobacteriophage Bxb1. *Mol Microbiol* 38, 955-970.
- Milani, C., Lugli, G.A., Duranti, S., Turrone, F., Bottacini, F., Mangifesta, M., Sanchez, B., Viappiani, A., Mancabelli, L., Taminiau, B., *et al.* (2014). Genomic encyclopedia of type strains of the genus *Bifidobacterium*. *Appl Environ Microbiol* 80, 6290-6302.
- Mobberley, J.M., Authement, R.N., Segall, A.M., and Paul, J.H. (2008). The temperate marine phage PhiHAP-1 of *Halomonas aquamarina* possesses a linear plasmid-like prophage genome. *J Virol* 82, 6618-6630.
- Monteiro, R., Pires, D.P., Costa, A.R., and Azeredo, J. (2018). Phage Therapy: Going Temperate? *Trends Microbiol.*
- Montgomery, M.T., Bustamante, C.A., Dedrick, R., Jacobs-Sera, D., and Hatfull, G.F. (2019). Yet more evidence of collusion: A new viral defense system encoded by *Gordonia* phage CarolAnn. *MBio*.
- Moreno Switt, A.I., Orsi, R.H., den Bakker, H.C., Vongkamjan, K., Altier, C., and Wiedmann, M. (2013). Genomic characterization provides new insight into *Salmonella* phage diversity. *BMC Genomics* 14, 481.
- Mrazek, J., and Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A* 95, 3720-3725.
- Nakashima, N., Mitani, Y., and Tamura, T. (2005). Actinomycetes as host cells for production of recombinant proteins. *Microb Cell Fact* 4, 7.
- Nesbit, C.E., Levin, M.E., Donnelly-Wu, M.K., and Hatfull, G.F. (1995). Transcriptional regulation of repressor synthesis in mycobacteriophage L5. *Mol Microbiol* 17, 1045-1056.
- Newton-Foot, M., and Gey van Pittius, N.C. (2013). The complex architecture of mycobacterial promoters. *Tuberculosis (Edinb)* 93, 60-74.
- Novick, R.P. (1987). Plasmid incompatibility. *Microbiol Rev* 51, 381-395.

- O'Callaghan, A., Bottacini, F., O'Connell Motherway, M., and van Sinderen, D. (2015). Pangenome analysis of *Bifidobacterium longum* and site-directed mutagenesis through bypass of restriction-modification systems. *BMC Genomics* *16*, 832.
- O'Connell Motherway, M., Zomer, A., Leahy, S.C., Reunanen, J., Bottacini, F., Claesson, M.J., O'Brien, F., Flynn, K., Casey, P.G., Munoz, J.A., *et al.* (2011). Functional genome analysis of *Bifidobacterium breve* UCC2003 reveals type IVb tight adherence (Tad) pili as an essential and conserved host-colonization factor. *Proc Natl Acad Sci U S A* *108*, 11217-11222.
- Ojha, A., Anand, M., Bhatt, A., Kremer, L., Jacobs, W.R., Jr., and Hatfull, G.F. (2005). GroEL1: a dedicated chaperone involved in mycolic acid biosynthesis during biofilm formation in mycobacteria. *Cell* *123*, 861-873.
- Oldfield, L.M., and Hatfull, G.F. (2014). Mutational analysis of the mycobacteriophage BPs promoter PR reveals context-dependent sequences for mycobacterial gene expression. *J Bacteriol* *196*, 3589-3597.
- Oliveira, J., Mahony, J., Hanemaaijer, L., Kouwen, T., Neve, H., MacSharry, J., and van Sinderen, D. (2017). Detecting *Lactococcus lactis* Prophages by Mitomycin C-Mediated Induction Coupled to Flow Cytometry Analysis. *Front Microbiol* *8*, 1343.
- Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* *17*, 132.
- Oppenheim, A.B., Kobiler, O., Stavans, J., Court, D.L., and Adhya, S. (2005). Switches in bacteriophage lambda development. *Annu Rev Genet* *39*, 409-429.
- Organization, W.H. (2013). Global Tuberculosis Report.
- Pabo, C.O., and Lewis, M. (1982). The operator-binding domain of lambda repressor: structure and DNA recognition. *Nature* *298*, 443-447.
- Pashley, C.A., Parish, T., McAdam, R.A., Duncan, K., and Stoker, N.G. (2003). Gene replacement in mycobacteria by using incompatible plasmids. *Appl Environ Microbiol* *69*, 517-523.
- Paul, J.H., Sullivan, M.B., Segall, A.M., and Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* *133*, 463-476.

- Payne, K.M., and Hatfull, G.F. (2012). Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities. *PLoS One* 7, e34052.
- Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* 113, 171-182.
- Pena, C.E., Lee, M.H., Pedulla, M.L., and Hatfull, G.F. (1997). Characterization of the mycobacteriophage L5 attachment site, attP. *J Mol Biol* 266, 76-92.
- Peters, J.M., Vangeloff, A.D., and Landick, R. (2011). Bacterial transcription terminators: the RNA 3'-end chronicles. *J Mol Biol* 412, 793-813.
- Petersen, J., Brinkmann, H., and Pradella, S. (2009). Diversity and evolution of repABC type plasmids in Rhodobacterales. *Environ Microbiol* 11, 2627-2638.
- Pham, T.T., Jacobs-Sera, D., Pedulla, M.L., Hendrix, R.W., and Hatfull, G.F. (2007). Comparative genomic analysis of mycobacteriophage Tweety: evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. *Microbiology* 153, 2711-2723.
- Pinto, U.M., Pappas, K.M., and Winans, S.C. (2012). The ABCs of plasmid replication and segregation. *Nat Rev Microbiol* 10, 755-765.
- Pitcher, R.S., Tonkin, L.M., Daley, J.M., Palmbo, P.L., Green, A.J., Velting, T.L., Brzostek, A., Korycka-Machala, M., Cresawn, S., Dziadek, J., *et al.* (2006). Mycobacteriophage exploit NHEJ to facilitate genome circularization. *Mol Cell* 23, 743-748.
- Popa, O., Landan, G., and Dagan, T. (2017). Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *ISME J* 11, 543-554.
- Pope, W.H., Bowman, C.A., Russell, D.A., Jacobs-Sera, D., Asai, D.J., Cresawn, S.G., Jacobs, W.R., Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., *et al.* (2015). Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* 4, e06416.
- Pope, W.H., Ferreira, C.M., Jacobs-Sera, D., Benjamin, R.C., Davis, A.J., DeJong, R.J., Elgin, S.C., Guilfoile, F.R., Forsyth, M.H., Harris, A.D., *et al.* (2011a). Cluster K mycobacteriophages: insights into the evolutionary origins of mycobacteriophage TM4. *PLoS One* 6, e26750.

- Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A., Alexander, L.M., Alfano, M.B., Alford, S.T., Amy, N.E., *et al.* (2011b). Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. *PLoS One* 6, e16329.
- Pope, W.H., Jacobs-Sera, D., Russell, D.A., Rubin, D.H., Kajee, A., Msibi, Z.N., Larsen, M.H., Jacobs, W.R., Jr., Lawrence, J.G., Hendrix, R.W., *et al.* (2014). Genomics and proteomics of mycobacteriophage patience, an accidental tourist in the *Mycobacterium* neighborhood. *MBio* 5, e02145.
- Pope, W.H., Mavrich, T.N., Garlena, R.A., Guerrero-Bustamante, C.A., Jacobs-Sera, D., Montgomery, M.T., Russell, D.A., Warner, M.H., Science Education Alliance-Phage Hunters Advancing, G., Evolutionary, S., *et al.* (2017). Bacteriophages of *Gordonia* spp. Display a Spectrum of Diversity and Genetic Relationships. *MBio* 8.
- Procopio, R.E., Silva, I.R., Martins, M.K., Azevedo, J.L., and Araujo, J.M. (2012). Antibiotics produced by *Streptomyces*. *Braz J Infect Dis* 16, 466-471.
- Proux, C., van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G.F., Desiere, F., and Brussow, H. (2002). The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J Bacteriol* 184, 6026-6036.
- Ptashne, M. (1992). A genetic switch : phage [lambda] and higher organisms, 2nd edn (Cambridge, Mass.: Cell Press : Blackwell Scientific Publications).
- Puigbo, P., Lobkovsky, A.E., Kristensen, D.M., Wolf, Y.I., and Koonin, E.V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* 12, 66.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842.
- Ranes, M.G., Rauzier, J., Lagranderie, M., Gheorghiu, M., and Gicquel, B. (1990). Functional analysis of pAL5000, a plasmid from *Mycobacterium fortuitum*: construction of a "mini" mycobacterium-*Escherichia coli* shuttle vector. *J Bacteriol* 172, 2793-2797.

- Ravin, N.V. (2015). Replication and Maintenance of Linear Phage-Plasmid N15. *Microbiol Spectr* 3, PLAS-0032-2014.
- Ravin, N.V., Svarchevsky, A.N., and Deho, G. (1999). The anti-immunity system of phage-plasmid N15: identification of the antirepressor gene and its control by a small processed RNA. *Mol Microbiol* 34, 980-994.
- Refardt, D. (2011). Within-host competition determines reproductive success of temperate bacteriophages. *ISME J* 5, 1451-1460.
- Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9, 173-175.
- Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci U S A* 110, 20236-20241.
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16, 276-277.
- Rohwer, F. (2003). Global phage diversity. *Cell* 113, 141.
- Rohwer, F., and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* 184, 4529-4535.
- Rohwer, F., and Thurber, R.V. (2009). Viruses manipulate the marine environment. *Nature* 459, 207-212.
- Rokney, A., Kobilier, O., Amir, A., Court, D.L., Stavans, J., Adhya, S., and Oppenheim, A.B. (2008). Host responses influence on the induction of lambda prophage. *Mol Microbiol* 68, 29-36.
- Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4.
- Roux, S., Hawley, A.K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S.J., and Sullivan, M.B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *Elife* 3, e03125.
- Rozanov, D.V., D'Ari, R., and Sineoky, S.P. (1998). RecA-independent pathways of lambdoid prophage induction in *Escherichia coli*. *J Bacteriol* 180, 6306-6315.

- Russell, D.A., and Hatfull, G.F. (2017). PhagesDB: the actinobacteriophage database. *Bioinformatics* 33, 784-786.
- Rybniiker, J., Plum, G., Robinson, N., Small, P.L., and Hartmann, P. (2008). Identification of three cytotoxic early proteins of mycobacteriophage L5 leading to growth inhibition in *Mycobacterium smegmatis*. *Microbiology* 154, 2304-2314.
- Sakaguchi, Y., Hayashi, T., Kurokawa, K., Nakayama, K., Oshima, K., Fujinaga, Y., Ohnishi, M., Ohtsubo, E., Hattori, M., and Oguma, K. (2005). The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc Natl Acad Sci U S A* 102, 17472-17477.
- Samson, J.E., Magadan, A.H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nat Rev Microbiol* 11, 675-687.
- Sandmeier, H. (1994). Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. *Mol Microbiol* 12, 343-350.
- Sandmeier, H., Iida, S., and Arber, W. (1992). DNA inversion regions Min of plasmid p15B and Cin of bacteriophage P1: evolution of bacteriophage tail fiber genes. *J Bacteriol* 174, 3936-3944.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687-695.
- Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162, 729-773.
- Santangelo, T.J., and Artsimovitch, I. (2011). Termination and antitermination: RNA polymerase runs a stop sign. *Nat Rev Microbiol* 9, 319-329.
- Sarkis, G.J., Jacobs, W.R., Jr., and Hatfull, G.F. (1995). L5 luciferase reporter mycobacteriophages: a sensitive tool for the detection and assay of live mycobacteria. *Mol Microbiol* 15, 1055-1067.
- Sau, K., Gupta, S.K., Sau, S., and Ghosh, T.C. (2005). Synonymous codon usage bias in 16 *Staphylococcus aureus* phages: implication in phage therapy. *Virus research* 113, 123-131.
- Sauer, B. (1987). Functional expression of the cre-lox site-specific recombination system in the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* 7, 2087-2096.

- Sauer, R.T., Krovatin, W., DeAnda, J., Youderian, P., and Susskind, M.M. (1983). Primary structure of the *immI* immunity region of bacteriophage P22. *J Mol Biol* 168, 699-713.
- Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M., and Pabo, C.O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature* 298, 447-451.
- Schicklmaier, P., and Schmieger, H. (1997). Sequence comparison of the genes for immunity, DNA replication, and cell lysis of the P22-related *Salmonella* phages ES18 and L. *Gene* 195, 93-100.
- Schooley, R.T., Biswas, B., Gill, J.J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J.J., Reed, S.L., Rohwer, F., Benler, S., *et al.* (2017). Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant *Acinetobacter baumannii* Infection. *Antimicrob Agents Chemother* 61.
- Schreiter, E.R., and Drennan, C.L. (2007). Ribbon-helix-helix transcription factors: variations on a theme. *Nat Rev Microbiol* 5, 710-720.
- Schumacher, M.A. (2012). Bacterial plasmid partition machinery: a minimalist approach to survival. *Curr Opin Struct Biol* 22, 72-79.
- Serebriiskii, I., Khazak, V., and Golemis, E.A. (1999). A two-hybrid dual bait system to discriminate specificity of protein interactions. *J Biol Chem* 274, 17080-17087.
- Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345-353.
- Shreiner, A.B., Kao, J.Y., and Young, V.B. (2015). The gut microbiome in health and in disease. *Curr Opin Gastroenterol* 31, 69-75.
- Sims, G.E., Jun, S.R., Wu, G.A., and Kim, S.H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A* 106, 2677-2682.
- Smith, H.W. (1972). Ampicillin resistance in *Escherichia coli* by phage infection. *Nat New Biol* 238, 205-206.
- Smith, M.C., Burns, R.N., Wilson, S.E., and Gregory, M.A. (1999). The complete genome sequence of the *Streptomyces* temperate phage straight phiC31: evolutionary relationships to other viruses. *Nucleic Acids Res* 27, 2145-2155.

- Smith, M.C., Hendrix, R.W., Dedrick, R., Mitchell, K., Ko, C.C., Russell, D., Bell, E., Gregory, M., Bibb, M.J., Pethick, F., *et al.* (2013). Evolutionary relationships among actinophages and a putative adaptation for growth in *Streptomyces* spp. *J Bacteriol* *195*, 4924-4935.
- Smith, M.C., and Owen, C.E. (1991). Three in-frame N-terminally different proteins are produced from the repressor locus of the *Streptomyces* bacteriophage phi C31. *Mol Microbiol* *5*, 2833-2844.
- Snapper, S.B., Melton, R.E., Mustafa, S., Kieser, T., and Jacobs, W.R., Jr. (1990). Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. *Mol Microbiol* *4*, 1911-1919.
- Sobel, J. (2005). Botulism. *Clin Infect Dis* *41*, 1167-1173.
- Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* *33*, W244-248.
- Speck, S.H., and Ganem, D. (2010). Viral latency and its regulation: lessons from the gamma-herpesviruses. *Cell Host Microbe* *8*, 100-115.
- Spinelli, S., Campanacci, V., Blangy, S., Moineau, S., Tegoni, M., and Cambillau, C. (2006). Modular structure of the receptor binding proteins of *Lactococcus lactis* phages. The RBP structure of the temperate phage TP901-1. *J Biol Chem* *281*, 14256-14262.
- Stayrook, S., Jaru-Ampornpan, P., Ni, J., Hochschild, A., and Lewis, M. (2008). Crystal structure of the lambda repressor and a model for pairwise cooperative operator binding. *Nature* *452*, 1022-1025.
- Stella, E.J., Franceschelli, J.J., Tasselli, S.E., and Morbidoni, H.R. (2013). Analysis of novel mycobacteriophages indicates the existence of different strategies for phage inheritance in mycobacteria. *PLoS One* *8*, e56384.
- Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* *22*, 1985-1994.
- Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. *Bioessays* *33*, 43-51.
- Stolt, P., and Stoker, N.G. (1996). Functional definition of regions necessary for replication and incompatibility in the *Mycobacterium fortuitum* plasmid pAL5000. *Microbiology* *142* (Pt 10), 2795-2802.

- Stolt, P., and Stoker, N.G. (1997). Mutational analysis of the regulatory region of the *Mycobacterium* plasmid pAL5000. *Nucleic Acids Res* 25, 3840-3846.
- Strohl, W.R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res* 20, 961-974.
- Subramanya, H.S., Arciszewska, L.K., Baker, R.A., Bird, L.E., Sherratt, D.J., and Wigley, D.B. (1997). Crystal structure of the site-specific recombinase, XerD. *EMBO J* 16, 5178-5187.
- Susskind, M.M., and Botstein, D. (1978). Molecular genetics of bacteriophage P22. *Microbiol Rev* 42, 385-413.
- Suttle, C.A. (2005). Viruses in the sea. *Nature* 437, 356-361.
- Swanson, W.J., and Vacquier, V.D. (1998). Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. *Science* 281, 710-712.
- Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 32, 1555-1556.
- Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. *Science* 278, 631-637.
- Thomas, J.M., Horspool, D., Brown, G., Tcherepanov, V., and Upton, C. (2007). GraphDNA: a Java program for graphical display of DNA composition analyses. *BMC Bioinformatics* 8, 21.
- Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14, 178-192.
- Tillier, E.R., and Collins, R.A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* 50, 249-257.
- Touchon, M., Bernheim, A., and Rocha, E.P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J* 10, 2744-2754.
- Tu, Q., and Lin, L. (2016). Gene content dissimilarity for subclassification of highly similar microbial strains. *BMC Genomics* 17, 647.
- Turroni, F., Peano, C., Pass, D.A., Foroni, E., Severgnini, M., Claesson, M.J., Kerr, C., Hourihane, J., Murray, D., Fuligni, F., *et al.* (2012). Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* 7, e36957.

- Twort, F.W. (1915). AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES. *The Lancet* 186, 1241-1243.
- Unniraman, S., Chatterji, M., and Nagaraja, V. (2002). DNA gyrase genes in *Mycobacterium tuberculosis*: a single operon driven by multiple promoters. *J Bacteriol* 184, 5449-5456.
- Utter, B., Deutsch, D.R., Schuch, R., Winer, B.Y., Verratti, K., Bishop-Lilly, K., Sozhamannan, S., and Fischetti, V.A. (2014). Beyond the chromosome: the prevalence of unique extra-chromosomal bacteriophages with integrated virulence genes in pathogenic *Staphylococcus aureus*. *PLoS One* 9, e100502.
- Valenzuela, D., and Ptashne, M. (1989). P22 repressor mutants deficient in co-operative binding and DNA loop formation. *EMBO J* 8, 4345-4350.
- Van Kaer, L., Gansemans, Y., Van Montagu, M., and Dhaese, P. (1988). Interaction of the *Bacillus subtilis* phage phi 105 repressor DNA: a genetic analysis. *EMBO J* 7, 859-866.
- Van Kaer, L., Van Montagu, M., and Dhaese, P. (1987). Transcriptional control in the EcoRI-F immunity region of *Bacillus subtilis* phage phi 105. Identification and unusual structure of the operator. *J Mol Biol* 197, 55-67.
- Van Kaer, L., Van Montagu, M., and Dhaese, P. (1989). Purification and in vitro DNA-binding specificity of the *Bacillus subtilis* phage phi 105 repressor. *J Biol Chem* 264, 14784-14791.
- van Kessel, J.C., and Hatfull, G.F. (2008). *Mycobacterial recombineering*. *Methods Mol Biol* 435, 203-215.
- Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., and Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 43, 6761-6771.
- Ventura, M., Canchaya, C., Bernini, V., Del Casale, A., Dellaglio, F., Neviani, E., Fitzgerald, G.F., and van Sinderen, D. (2005a). Genetic characterization of the *Bifidobacterium breve* UCC 2003 hrcA locus. *Appl Environ Microbiol* 71, 8998-9007.
- Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G.F., Chater, K.F., and van Sinderen, D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* 71, 495-548.
- Ventura, M., Lee, J.H., Canchaya, C., Zink, R., Leahy, S., Moreno-Munoz, J.A., O'Connell-Motherway, M., Higgins, D., Fitzgerald, G.F., O'Sullivan, D.J., *et al.* (2005b). Prophage-like

- elements in bifidobacteria: insights from genomics, transcription, integration, distribution, and phylogenetic analysis. *Appl Environ Microbiol* *71*, 8692-8705.
- Ventura, M., Turrioni, F., Lima-Mendez, G., Foroni, E., Zomer, A., Duranti, S., Giubellini, V., Bottacini, F., Horvath, P., Barrangou, R., *et al.* (2009). Comparative analyses of prophage-like elements present in bifidobacterial genomes. *Appl Environ Microbiol* *75*, 6929-6936.
- Villanueva, V.M., Oldfield, L.M., and Hatfull, G.F. (2015). An Unusual Phage Repressor Encoded by Mycobacteriophage BPs. *PLoS One* *10*, e0137187.
- Waagmeester, A., Thompson, J., and Reyrat, J.M. (2005). Identifying sigma factors in *Mycobacterium smegmatis* by comparative genomic analysis. *Trends Microbiol* *13*, 505-509.
- Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* *272*, 1910-1914.
- Wang, S., Qiao, X., Liu, X., Zhang, X., Wang, C., Zhao, X., Chen, Z., Wen, Y., and Song, Y. (2010a). Complete genomic sequence analysis of the temperate bacteriophage phiSASD1 of *Streptomyces avermitilis*. *Virology* *403*, 78-84.
- Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M., and Wood, T.K. (2010b). Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* *1*, 147.
- Wang, X., Montero Llopis, P., and Rudner, D.Z. (2013). Organization and segregation of bacterial chromosomes. *Nat Rev Genet* *14*, 191-203.
- Wang, X., and Wood, T.K. (2016). Cryptic prophages as targets for drug development. *Drug Resist Updat* *27*, 30-38.
- Weigel, C., and Seitz, H. (2006). Bacteriophage replication modules. *FEMS Microbiol Rev* *30*, 321-381.
- Weinacht, K.G., Roche, H., Krinos, C.M., Coyne, M.J., Parkhill, J., and Comstock, L.E. (2004). Tyrosine site-specific recombinases mediate DNA inversions affecting the expression of outer surface proteins of *Bacteroides fragilis*. *Mol Microbiol* *53*, 1319-1330.
- Wilson, S.E., Ingham, C.J., Hunter, I.S., and Smith, M.C. (1995). Control of lytic development in the *Streptomyces* temperate phage phi C31. *Mol Microbiol* *16*, 131-143.
- Wolf, Y.I., Makarova, K.S., Lobkovsky, A.E., and Koonin, E.V. (2016). Two fundamentally different classes of microbial genes. *Nat Microbiol* *2*, 16208.

- Wood, H.E., Devine, K.M., and McConnell, D.J. (1990). Characterisation of a repressor gene (xre) and a temperature-sensitive allele from the *Bacillus subtilis* prophage, PBSX. *Gene* *96*, 83-88.
- Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H., and Ussery, D.W. (2006). Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* *8*, 353-361.
- Yang, L., Nielsen, A.A., Fernandez-Rodriguez, J., McClune, C.J., Laub, M.T., Lu, T.K., and Voigt, C.A. (2014). Permanent genetic memory with >1-byte capacity. *Nat Methods* *11*, 1261-1266.
- Yarmolinsky, M.B. (2004). Bacteriophage P1 in retrospect and in prospect. *J Bacteriol* *186*, 7025-7028.
- Youssef, N.H., Couger, M.B., McCully, A.L., Criado, A.E., and Elshahed, M.S. (2015). Assessing the global phylum level diversity within the bacterial domain: A review. *J Adv Res* *6*, 269-282.
- Zabala, B., Hammerl, J.A., Espejo, R.T., and Hertwig, S. (2009). The linear plasmid prophage Vp58.5 of *Vibrio parahaemolyticus* is closely related to the integrating phage VHML and constitutes a new incompatibility group of telomere phages. *J Virol* *83*, 9313-9320.
- Zampini, M., Derome, A., Bailey, S.E., Barilla, D., and Hayes, F. (2009). Recruitment of the ParG segregation protein to different affinity DNA sites. *J Bacteriol* *191*, 3832-3841.
- Zhang, C.T., Zhang, R., and Ou, H.Y. (2003). The Z curve database: a graphic representation of genome sequences. *Bioinformatics* *19*, 593-599.
- Zhang, H., Gao, S., Lercher, M.J., Hu, S., and Chen, W.H. (2012a). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res* *40*, W569-572.
- Zhang, Z., Liu, Y., Wang, S., Yang, D., Cheng, Y., Hu, J., Chen, J., Mei, Y., Shen, P., Bamford, D.H., *et al.* (2012b). Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205. *Virology* *434*, 233-241.
- Zhong, L., Cheng, Q., Tian, X., Zhao, L., and Qin, Z. (2010). Characterization of the replication, transfer, and plasmid/lytic phage cycle of the *Streptomyces* plasmid-phage pZL12. *J Bacteriol* *192*, 3747-3754.
- Zhu, W., Wang, J., Zhu, Y., Tang, B., Zhang, Y., He, P., Zhang, Y., Liu, B., Guo, X., Zhao, G., *et al.* (2015). Identification of three extra-chromosomal replicons in *Leptospira* pathogenic strain and development of new shuttle vectors. *BMC Genomics* *16*, 90.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res* 37, e151.